

STOCHASTIC OPTIMIZATION ALGORITHMS IN MACHINE LEARNING: DYNAMICS, CONVERGENCE, GENERALIZATION

WAÏSS AZIZIAN

ADVISORS

FRANCK IUTZELER	Université de Toulouse
JÉRÔME MALICK	CNRS, Université Grenoble Alpes
PANAYOTIS MERTIKOPOULOS	CNRS, Université Grenoble Alpes

REVIEWERS

FRANCIS BACH	Inria, École Normale Supérieure
JOHN DUCHI	Stanford University

EXAMINERS

NIAO HE	ETH Zürich
ANATOLI JUDITSKY	Université Grenoble Alpes
JULIEN MAIRAL	Inria, Université Grenoble Alpes
GABRIEL PEYRÉ	CNRS, École Normale Supérieure



Université Grenoble Alpes
École Doctorale MSTII
Spécialité : Mathématiques Appliquées
Defended on June 16, 2026

ABSTRACT

This thesis considers machine learning through the lens of mathematical optimization, and is concerned with several issues that can arise: non-convexity of the loss landscape, uncertainty and shifts in the data, and intricate structure of the parameter space. In the first part of this thesis, we focus on stochastic gradient methods, and in particular stochastic gradient descent (SGD), which remains the workhorse for training today's largest machine learning systems. Despite SGD's long history and overwhelming success, one fundamental question about its behavior on general objectives remains open: what is the long-run behavior of SGD on non-convex landscapes? In particular, which regions of the parameter space are visited the most by SGD? By how much? And how long does it take for SGD to reach a set of parameters of interest, for instance the global minimum? In this thesis, our goal is to give a general, principled answer for arbitrary non-convex objectives. Using large deviations and the theory of randomly perturbed dynamical systems, we show that the long-run distribution of SGD takes the form of a Boltzmann–Gibbs law: the step-size acts as a temperature, and the "energy levels" are determined jointly by the objective and the noise. As a consequence, SGD concentrates exponentially around the minimum-energy state — which need not be the global minimum — while other regions are visited with probabilities exponentially proportional to their energy. The same tools yield matching upper and lower bounds on the time SGD takes to reach a neighborhood of the global minimum, showing that this time is controlled by the most "costly" obstacles in the landscape, linking the problem's global geometry with the statistics of the noise. Taken together, these works provide a sharp characterization of the long-run behavior of SGD on general non-convex objectives.

The second part of this thesis turns to robust and constrained optimization. We first study Wasserstein distributionally robust optimization (WDRO), a framework that hedges against distribution shifts by minimizing the worst-case loss over a Wasserstein ball around the empirical distribution. We propose an entropic regularization framework that makes the formulation tractable in standard machine learning processes, and we establish the first exact generalization bounds — upper bounds on the true risk without error terms — that avoid the curse of dimensionality in the choice of the radius of the ball. We then move to constrained optimization and analyze the last-iterate convergence of Bregman proximal methods on constrained variational inequalities, introducing the *Legendre exponent* — a local regularity measure of the Bregman function near a solution. We show that this quantity determines the convergence rate in both deterministic and stochastic settings, yielding a strict linear/sublinear dichotomy depending on the choice of regularizer.

RÉSUMÉ

Cette thèse aborde l'apprentissage automatique sous l'angle de l'optimisation mathématique, et s'intéresse à plusieurs difficultés qui peuvent survenir : la non-convexité de la fonction objectif, l'incertitude et les changements de distribution des données, ainsi que la structure complexe de l'espace des paramètres. La première partie de cette thèse porte sur les méthodes de gradient stochastique, et en particulier la descente de gradient stochastique (SGD), qui reste le pilier de l'entraînement des grands systèmes d'apprentissage automatique actuels. Malgré la longue histoire et le succès considérable de SGD, une question fondamentale concernant son comportement sur des objectifs génériques reste ouverte : quel est le comportement à long terme de SGD sur des objectifs non convexes ? En particulier, quelles régions de l'espace des paramètres sont les plus visitées par SGD ? Dans quelle mesure ? Et combien de temps faut-il à SGD pour atteindre un ensemble de paramètres d'intérêt, par exemple le minimum global ? Dans cette thèse, notre objectif est de fournir une réponse générale et mathématiquement fondée pour des objectifs non convexes arbitraires. En s'appuyant sur la théorie des grandes déviations et la théorie des systèmes dynamiques aléatoirement perturbés, nous montrons que la distribution à long terme de SGD prend la forme d'une loi de Boltzmann–Gibbs : le pas joue le rôle d'une température, et les « niveaux d'énergie » sont déterminés conjointement par l'objectif et le bruit. Par conséquent, SGD se concentre exponentiellement autour de l'état d'énergie minimale — qui n'est pas nécessairement le minimum global — tandis que les autres régions sont visitées avec des probabilités exponentiellement proportionnelles à leur énergie. Les mêmes outils nous permettent d'obtenir des bornes supérieures et inférieures correspondantes sur le temps que met SGD pour atteindre un voisinage du minimum global, montrant que ce temps est contrôlé par les obstacles les plus « coûteux » du paysage de la fonction de perte, reliant la géométrie globale du problème aux statistiques du bruit. Ces travaux fournissent donc une caractérisation fine du comportement à long terme de SGD sur des objectifs non convexes généraux.

La deuxième partie de cette thèse se tourne vers l'optimisation robuste et sous contraintes. Nous étudions d'abord l'optimisation distributionnellement robuste au sens de Wasserstein (WDRO), un cadre qui se prémunit contre les changements de distribution en minimisant la perte dans le pire cas sur une boule de Wasserstein autour de la distribution empirique. Nous proposons un cadre de régularisation entropique qui rend la formulation tractable dans les processus standards d'apprentissage automatique, et nous établissons les premières bornes de généralisation exactes — des bornes supérieures sur le risque réel sans termes d'erreur — qui évitent le fléau de la dimension dans le choix du rayon de la boule. Nous abordons ensuite l'optimisation sous contraintes et analysons la convergence du dernier itéré des méthodes proximales de Bregman sur des inégalités variationnelles sous contraintes, en introduisant l'*exposant de Legendre* — une mesure de régularité locale de la fonction de Bregman au voisinage d'une solution. Nous montrons que cette quantité détermine le taux de convergence dans les cadres déterministe et stochastique, donnant lieu à une dichotomie stricte linéaire/sous-linéaire selon le choix du régulariseur.

ACKNOWLEDGMENTS

My first thanks go to my advisors Jérôme Malick, Franck Iutzeler and Panayotis Mertikopoulos, for both the guidance and the freedom they gave me during these years. The PhD has been an extremely enriching experience for me, and it is largely thanks to you. Thank you for the freedom you gave me, for believing in me even when nothing worked, and for the advice when I needed it. Thank you for all the moments we spent together, from the meals at Martins to the very short nights before deadlines. To me, you are an incredible supervision team, and I consider myself extremely lucky to have worked with you.

I would like to thank the jury for agreeing to take part in this defense. I am honored that you all accepted. Thank you to Niao He for your interest in our work and your kind words. Thank you to Julien Mairal for agreeing to be president of this jury and for inviting me several times to his team, which made me feel welcome in the Grenoble research community. Thank you to Anatoli Juditsky for the insightful comments and discussions over the years, in particular when we were working on Bregman algorithms. Thank you to Gabriel Peyré for following my work and providing guidance all these years, first at ENS and then as a CST expert. Thank you to Francis Bach and John Duchi for agreeing to review the manuscript. Thank you, John, for the discussions already back in Les Houches a year ago. Thank you, Francis, for the guidance across the years, as well as for giving the ML class at ENS where I first learned about SGD.

I would like to thank everybody with whom I collaborated before and during the PhD. In particular, I thank Simon Lacoste-Julien and Gauthier Gidel for introducing me to optimization research, as well as Marc Lelarge. I would also like to thank Ali Hasan for the great experience last summer, I learned a lot.

More generally, I would like to thank everyone at LJK: the whole administrative team and all the permanent staff, who contribute to the scientific environment of the lab. In particular, thank you to all the non-permanent members, both those who left several years ago and those who have only just arrived, who make LJK a lab where it is so good to do research.

I would especially like to thank the former DAO students who welcomed me to Grenoble: Yassine, Yu-Gan, and Gilles. Thank you to Gilles, Alice, and Thibault for introducing me to hiking, in particular to some slightly dubious hikes... Thank you to those who introduced me to climbing, in particular Anatole and Sergei (as well as the physiotherapist for fixing me afterwards), and to the other LJK climbers. Thank you to those who were almost our office mates, Yannis and Sasilla, for the drinks on the terrace and the laughter. Finally, my office mates: Pierre-Louis, Florian, Victor, and Hadi. Thank you, Pierre-Louis, for the scientific discussions and the topological struggles. We went through all kinds of moments in this office and outside it. Florian, Victor: a lot happened during these years, but they were great years. Once again, thank you to all the

non-permanent members, as well as those from Inria, in particular Julien Zhou; I cannot name everyone.

Thank you to my friends from school, in particular Manu, thanks to whom I survived *classes préparatoires* and ENS. Finally, thank you to my family, who have always supported and encouraged me, and believed in me. And finally, Ieva, without whom I simply could not have done this PhD. Thank you all.

CONTENTS

ACKNOWLEDGMENTS vii

1 INTRODUCTION 1

- 1.1 Context: Empirical Risk Minimization 1
- 1.2 Thread 1: Dynamics of SGD in non-convex problems 2
- 1.3 Thread 2: Robust Optimization with Wasserstein Uncertainty 3
- 1.4 Thread 3: Constrained Variational Inequalities 4
- 1.5 Outline of the Thesis 5
- 1.6 Publications 6

PART I ON THE LONG-RUN BEHAVIOR OF SGD 9

2 OVERVIEW 11

- 2.1 Introduction 11
- 2.2 Overview of our results 12
- 2.3 Related Work 16

3 THE LARGE DEVIATIONS VIEWPOINT 21

- 3.1 Large deviations, long-run behavior, and gradient flow 22
- 3.2 Setup 25
- 3.3 Large deviations for SGD 29
- 3.4 The induced chain 35
- 3.5 Reduction to a finite state-space 36
- 3.6 Transition times 41
- 3.7 Transition probabilities 45
- 3.8 Summary 48

4 THE INVARIANT MEASURE OF SGD 51

- 4.1 Characterizing the invariant measure 51
- 4.2 Applications and implications 59

5 THE GLOBAL CONVERGENCE TIME OF SGD 71

- 5.1 An instructive example 71
- 5.2 Hitting-time analysis 75
- 5.3 Extensions and refinements 85

PART II DISTRIBUTIONALLY ROBUST AND CONSTRAINED MIN-MAX OPTIMIZATION 93

6 WASSERSTEIN DISTRIBUTIONALLY ROBUST OPTIMIZATION 95

- 6.1 Introduction 95
- 6.2 Setup and Assumptions 97
- 6.3 Main results and discussions 99
- 6.4 Regularized WDRO 103
- 6.5 Conclusion 109

7 THE RATE OF CONVERGENCE OF BREGMAN PROXIMAL METHODS 113

- 7.1 Introduction 113
- 7.2 Problem setup and preliminaries 115
- 7.3 Motivating examples 118
- 7.4 The Legendre exponent and convergence rate analysis 120

7.5	Finer results for linearly constrained problems	129
7.6	Mirror descent on stochastic variational inequalities	138
7.7	Conclusion	144
8	PERSPECTIVES	145
8.1	Towards a principled understanding of stochastic optimization in deep learning	145
8.2	Building machine learning systems that we can trust	147
	BIBLIOGRAPHY	149
	APPENDICES	167
A	WHAT IS THE LONG-RUN DISTRIBUTION OF SGD?	169
B	THE GLOBAL CONVERGENCE TIME OF SGD	229
C	REGULARIZATION FOR WDRO	281
D	EXACT GENERALIZATION GUARANTEES FOR (REGULARIZED) WDRO	313
E	THE RATE OF CONVERGENCE OF BREGMAN PROXIMAL METHODS	357
F	LAST-ITERATE CONVERGENCE OF OPTIMISTIC MIRROR DESCENT	363
G	LIST OF PUBLICATIONS	381

INTRODUCTION

1.1	Context: Empirical Risk Minimization	1
1.2	Thread 1: Dynamics of SGD in non-convex problems	2
1.3	Thread 2: Robust Optimization with Wasserstein Uncertainty	3
1.4	Thread 3: Constrained Variational Inequalities	4
1.5	Outline of the Thesis	5
1.6	Publications	6

Machine learning is, at its core, an optimization problem: finding a model that best describes the data at hand. This thesis thus considers machine learning through the lens of mathematical optimization.

In this introductory chapter, we outline our main contributions and the structure of the thesis. A detailed presentation of our results, along with a review of the relevant literature, is provided in the subsequent chapters.

1.1 CONTEXT: EMPIRICAL RISK MINIMIZATION

Machine learning (ML) algorithms aim to find a parametrized model that accurately represents the data at hand, and they achieve this by solving an empirical risk minimization (ERM) problem. Typically, this is formulated as follows: given n data points $\xi_1, \dots, \xi_n$ coming from some distribution $\mathbb{P}_{\text{data}}$, the associated ERM consists in finding the parameters of the model θ that solve

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n f(\theta, \xi_i) \quad (\text{ERM})$$

where $f(\theta, \xi)$ is the loss of the model on data point ξ . This formulation is the bedrock on which ML is built: it underpins learning in the vast majority of models, from linear regression to large language models.

However, this comes with several fundamental challenges:

- Gradient-based methods remain the algorithms of choice to solve (ERM), even though their convergence guarantees are largely confined to the convex setting. However, in neural networks and deep learning, $f(\theta, \xi)$ is typically (highly) non-convex in θ .
- When solving (ERM), the obtained parameters $\hat{\theta}$ tend to perform well relative to the training data $\xi_1, \dots, \xi_n$. However, because of overfitting or distribution shift, this may fail to be the case for new data points $\xi \sim \mathbb{P}_{\text{data}}$ that were not part of the training set, or for test points $\xi \sim \mathbb{P}_{\text{test}}$ which differ from $\mathbb{P}_{\text{data}}$.
- The parameter θ may be restricted to lie in a constraint set Θ (e.g., a polyhedron), in which case specific methods are required for training.

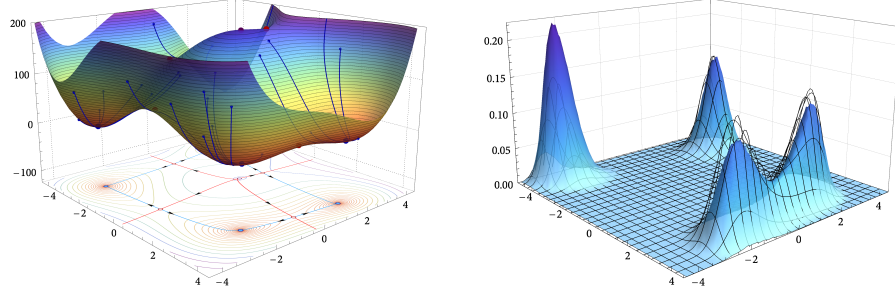


Figure 1: Graphical illustration of the invariant measure of SGD for the Himmelblau test function. The figure to the left depicts the loss landscape of f with several (deterministic) gradient flow orbits superimposed for visual convenience. The figure to the right illustrates the long-run distribution of 1000 samples of SGD run with a fixed step-size. the density landscape represents the observed distribution, while the superimposed wireframe indicates our theoretical prediction.

The goal of this thesis is to provide some partial answers to these challenges, organized into three main threads: the dynamics of stochastic gradient methods in non-convex problems, robust optimization with Wasserstein uncertainty sets, and constrained variational inequalities. We provide a quick overview of each of these threads below.

1.2 THREAD 1: DYNAMICS OF SGD IN NON-CONVEX PROBLEMS

Some of the most striking successes of deep learning, from large language models to reinforcement learning, share the same fundamental backbone: their training algorithms (which are stochastic first-order optimization methods). Nonetheless, the inner workings of these methods remain fairly opaque: our theoretical understanding mainly stems from the convex regime, but ERM objectives are often highly non-convex. The first (and most extensive) part of this thesis aims to shed some light on this disconnect through a novel framework for analyzing the long-run behavior of stochastic gradient descent (SGD), arguably the most central optimization method in ML. In more detail, we focus on the following questions:

► WHERE ARE THE ITERATES OF SGD MOST LIKELY TO BE OBSERVED IN THE LONG RUN?

We provide a partial answer to this crucial question via a crisp characterization of the invariant measures of SGD in non-convex problems. More precisely, we show that the long-run distribution of SGD takes the form of a Boltzmann–Gibbs law: the probability of SGD being close to a set of critical points $\mathcal{K}$ is of the form $\exp(-V(\mathcal{K})/\eta)$, where η is the method’s step-size and $V(\mathcal{K})$ is an “energy level” determined by f and the statistics of the gradient noise; see [Figure 1](#) for an illustration. Concretely, this means that SGD concentrates exponentially more mass in the vicinity of the problem’s ground state (the minimum of the energy), while all other regions are visited with probability exponentially smaller. This characterization of the long-run distribution of SGD opens the door to a deeper understanding of the optimization dynamics of modern machine learning systems and, in turn, quantitative predictions on their generalization, robustness and privacy properties

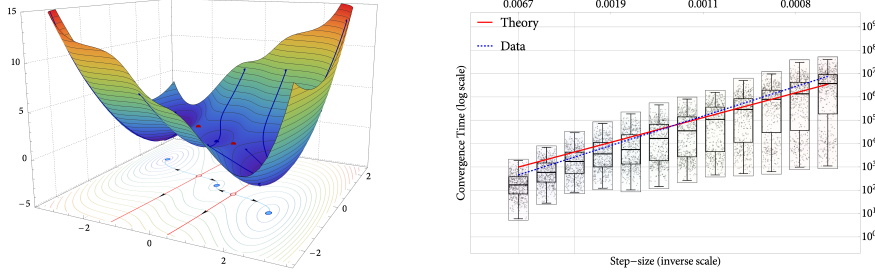


Figure 2: Visualization of the theoretical results on the convergence time of SGD to the global minimum for the three-hump camel function. The figure to the left depicts the loss landscape of f with several (deterministic) gradient flow orbits superimposed for visual convenience. The figure to the right shows the statistical analysis of the number of iterations required for SGD to reach the global minimum. For each value of the step-size, we performed 500 runs and recorded the number of iterations required to reach the vicinity of the global minimum. As predicted by our results, we observe that the global convergence time scales exponentially with the inverse of the step-size, with strong agreement between theory and experiment.

► HOW MUCH TIME DOES SGD TAKE TO REACH A GLOBAL MINIMUM?

We quantify the time stochastic gradient descent (SGD) takes to reach a neighborhood of a global minimizer on non-convex problems. Specifically, we show that the expected global convergence time of SGD started at x with constant step-size η scales (tightly) as $\exp(E(x)/\eta)$, where $E(x)$ is a geometric measure of the problem's hardness: the higher $E(x)$, the longer it takes for SGD to reach the global minimum. See Figure 2 for an illustration of this result. More precisely, $E(x)$ can be bounded by a quantity proportional to the number of spurious local minima and the depth of the deepest spurious minimum. In particular, $E(x)$ is small when the landscape is favorable, for instance when there is a small number of relatively shallow local minimizers. Conversely, $E(x)$ is large when the landscape is rough, with many deep spurious minima standing between the algorithm's starting point and the global minimum of the objective.

1.3 THREAD 2: ROBUST OPTIMIZATION WITH WASSERSTEIN UNCERTAINTY

Wasserstein distributionally robust optimization (WDRO) has recently emerged as a powerful framework for handling uncertainty in the data distribution. In standard empirical risk minimization (ERM), the model is trained to minimize the expected loss under the empirical distribution $\hat{\mathbb{P}}_n$, which assigns equal mass to the observed data points. However, this approach can be sensitive and fragile relative to discrepancies between the empirical and true data distributions. WDRO addresses this limitation by instead minimizing the worst-case expected loss over all probability distributions that lie within a specified Wasserstein distance of $\hat{\mathbb{P}}_n$, thereby explicitly accounting for potential distributional shifts (see Figure 3).

► ENTROPIC REGULARIZATION IN WDRO

Though WDRO has already been successfully applied in various contexts, this comes with significant drawbacks: it remains computationally challenging and leads to poorly conditioned objectives. To address these issues, we propose to regularize the inner maximization problem of WDRO with an entropic penalty inspired by the success of entropic regularization in optimal transport. This

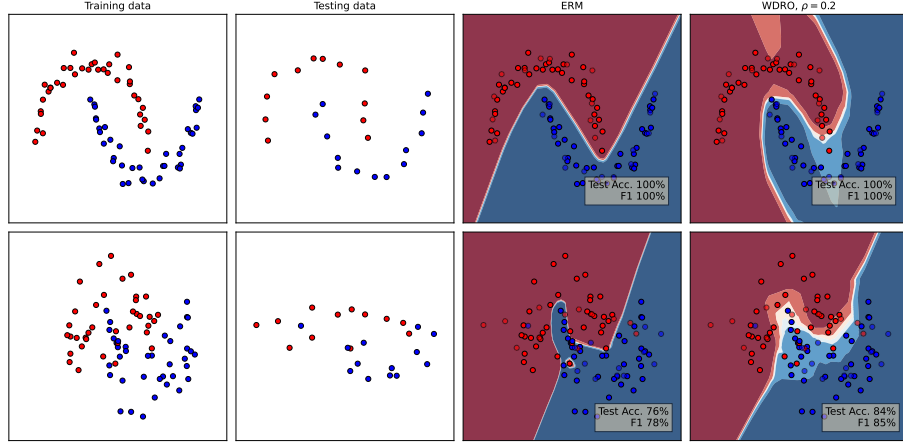


Figure 3: Two illustrations of the behaviour of (regularized) WDRO to train a simple multilayer perceptron on classification problems (two moons and a Gaussian-mixture classification task). The plots compare models trained with ERM and with regularized WDRO and come from the documentation of [Vincent et al. \(2026\)](#).

leads to a smoothed formulation that is easier to implement in existing machine learning contexts.

► GENERALIZATION AND ROBUSTNESS GUARANTEES

From a modeling standpoint, one of the key appeals of WDRO is its strong generalization guarantees: if the samples from $\hat{\mathbb{P}}_n$ are drawn from some true distribution $\mathbb{P}_{\text{data}}$, then the WDRO objective provides an exact upper-bound with high probability on the true risk, i.e., the expected loss over $\mathbb{P}_{\text{data}}$, as soon as the number of samples n is large enough. However, this approach suffers from the curse of dimensionality as it relies on non-parametric measure concentration: the required number of samples to ensure this guarantee scales exponentially with the dimension of the data. We provide new generalization bounds for WDRO that do not suffer from the curse of dimensionality and we showed that the WDRO training objective provides an exact upper bound on the true risk for general classes of models.

1.4 THREAD 3: CONSTRAINED VARIATIONAL INEQUALITIES

In this line of work, we focus on the study of methods designed for constrained problems and in particular Bregman gradient methods: by changing the geometry of the space, these methods can be used to solve constrained problems without the need for projections and often with attractive convergence properties. However, most of the existing convergence guarantees for these methods concern the average of the iterates. The question of the convergence of the last iterate of Bregman gradient methods has been open for a long time, even in the deterministic case.

To treat this question in as general a manner as possible, we focus on variational inequality (VI) problems of the form

$$\text{Find } x^* \in \mathcal{X} \text{ such that } \langle v(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{X}, \quad (\text{VI})$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a convex set and $v: \mathcal{X} \rightarrow \mathbb{R}^d$ is a vector field. This problem is a staple of many areas of mathematical programming, game theory and data

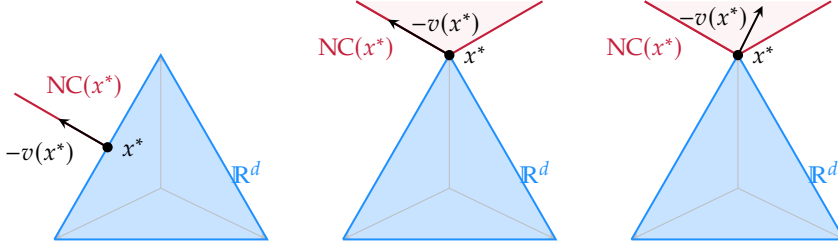


Figure 4: Different boundary solution configurations on the 2-dimensional unit simplex $\mathbb{R}^3$: each configuration leads to a different convergence regime.

science, as it provides a template for “optimization beyond minimization” – i.e., for problems where finding an optimal solution does not necessarily involve minimizing a loss function.

► **DETERMINISTIC CASE** We obtain a crisp characterization of last-iterate convergence rate of a wide range of Bregman gradient methods in terms of the local geometry induced by the underlying Bregman function near a given solution of (VI). To make this dependence precise, we introduce the notion of the *Legendre exponent*, a regularity measure for Bregman methods, and we show how the convergence rate of these methods is determined by this exponent (see Figures 4 and 5).

► **STOCHASTIC CASE** We then extend some of these results to the stochastic case, where the operator v is only accessible only through a noisy oracle. Again, the algorithm’s rate of convergence depends sharply on the local geometry induced by the underlying Bregman function and this dependence can be made precise via the Legendre exponent. The optimal step-size schedules, remarkably, also depend on the Legendre exponent, and they differ from the ones that are optimal for the ergodic convergence of these methods.

1.5 OUTLINE OF THE THESIS

To streamline our presentation, the thesis is structured in two parts, as follows:

Part I: On the Long-run Behavior of SGD: A Large Deviations Analysis.

The first part of this thesis revisits the ICML’24 and ICML’25 papers [Azizian et al. \(2024c\)](#) and [Azizian et al. \(2025a\)](#) to put the original approach, the results and the mathematics in a broader perspective.

[Chapter 2](#) introduces the two central questions studied in this part: characterizing the long-run distribution of the iterates of SGD and estimating the global convergence time of the method.

[Chapter 3](#) presents a detailed sketch of proof common to our results. It introduces the main tools — induced Markov chains, large deviation estimates — and outlines the proof strategy shared by these two papers.

[Chapter 4](#) gives the complete large-deviation characterization of the invariant measures of SGD on non-convex objectives, corresponding to the contributions of [Azizian et al. \(2024b\)](#).

[Chapter 5](#) establishes sharp estimates on the global convergence time of SGD via large deviations, corresponding to the contributions of [Azizian et al. \(2025a\)](#).

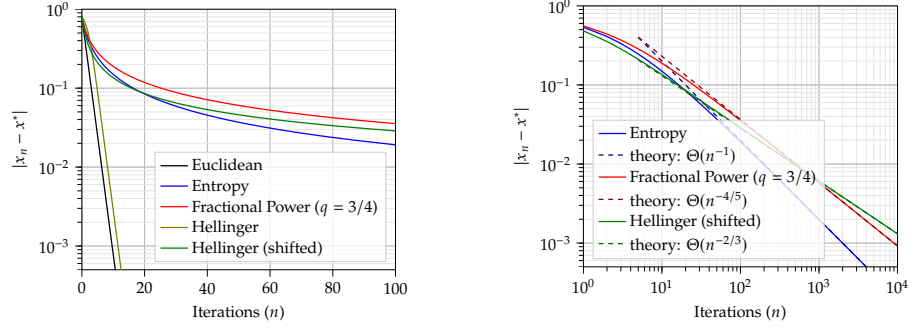


Figure 5: For a fixed simple (VI), convergence curves of Mirror Descent for different Bregman functions with different Legendre exponents (left: 100 iterations, right: 10^4 iterations). We observe different behaviors: linear convergence for Euclidean and shifted Hellinger and sublinear convergence with different rates for the other examples.

Part II: Distributionally Robust and Constrained Min-Max Optimization.

This second part presents the contributions of the two topics of § 1.3 and § 1.4.

Chapter 6 studies Wasserstein distributionally robust optimization (WDRO). It covers the exact generalization guarantees of the NeurIPS’23 paper (Azizian et al., 2023c) and the regularization framework of the ESAIM: COCV’23 article Azizian et al. (2023a).

Chapter 7 studies the last-iterate convergence of Bregman proximal methods as a function of the local geometry of the prox-mapping, establishing a sharp linear/sublinear dichotomy governed by the Legendre exponent of the Bregman function. It is based on the SIOPT article (Azizian et al., 2024a), with an extension to the stochastic setting from the COLT’21 paper (Azizian et al., 2021a).

Much of this manuscript has been written in a more conversational style, placing less emphasis on the precise formalism of the various notions, and more on the intuition and ideas driving the development of the subject matter. This choice is deliberate and reflects a conscious effort on the part of the author to make the material more accessible to non-experts in the field, and to present a coherent narrative that the rest of the community can build on. The expert reader who is interested in the (many) technical details and difficulties involved is deferred to the series of technical appendices (and original papers) referenced in the main text.

1.6 PUBLICATIONS

This manuscript includes original contributions by the author from the papers listed below. Further details can be found at the beginning of each section; a complete list of publications is provided in [Appendix G](#).

- **What is the long-run distribution of stochastic gradient descent? A large deviations analysis.**

ICML, 2024. (Azizian et al., 2024b)

- **The global convergence time of stochastic gradient descent in non-convex landscapes: Sharp estimates via large deviations.**

ICML, 2025. (Azizian et al., 2025a)

- **Exact generalization guarantees for (regularized) Wasserstein distributionally robust models.**

NeurIPS, 2023. ([Azizian et al., 2023c](#))

- **Regularization for Wasserstein distributionally robust optimization.**

ESAIM: COCV, 2023. ([Azizian et al., 2023a](#))

- **The rate of convergence of Bregman proximal methods: Local geometry vs. regularity vs. sharpness.**

SIAM Journal on Optimization, 2024. ([Azizian et al., 2024a](#))

- **The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities.**

COLT, 2021. ([Azizian et al., 2021a](#))

Besides my main work, I had the chance to collaborate with other researchers. Some side projects led to publications or preprints:

- **Skwdro: A library for Wasserstein distributionally robust machine learning.**

Vincent Florian, Waïss Azizian, Franck Iutzeler, and Jérôme Malick.

Journal of Machine Learning Research, 2026. ([Vincent et al., 2026](#))

This paper presents skwdro, a Python library for training distributionally robust machine learning models based on optimal-transport Wasserstein distances, with a PyTorch wrapper and scikit-learn-compatible estimators.

- **How does the pretraining distribution shape in-context learning? Task selection, generalization, and robustness.**

Waïss Azizian and Ali Hasan.

arXiv preprint arXiv:2510.01163, 2025. ([Azizian and Hasan, 2025](#))

We characterize theoretically how the pretraining distribution determines task selection, generalization, and robustness in in-context learning.

- **Almost sure convergence of stochastic gradient methods under gradient domination.**

Simon Weissmann, Sara Klein, Waïss Azizian, and Leif Döring.

Transactions on Machine Learning Research, 2025. ([Weissmann et al., 2025](#))

We prove almost sure convergence rates for the last iterate of SGD under gradient domination assumptions, with rates matching recent results in expectation up to an arbitrarily small gap.

- **The geometries of truth are orthogonal across tasks.**

Waïss Azizian, Michael Kirchhof, Eugene Ndiaye, Louis Bethune, Michal Klein, Pierre Ablin, and Marco Cuturi.

ICML 2025 Workshop on Reliable and Responsible Foundation Models, 2025. ([Azizian et al., 2025b](#))

We show that “geometries of truth” learned from model activations are intrinsically task-dependent and fail to transfer across tasks, and that richer probing strategies do not remove this limitation.

The following publications predate the start of my PhD.

- **Expressive power of invariant and equivariant graph neural networks.**

Waïss Azizian and Marc Lelarge.

ICLR, 2021. (Azizian and Lelarge, 2021)

We propose a theoretical framework to compare the expressive power of graph neural networks (GNNs) and prove the first approximation guarantees for practical invariant and equivariant graph neural networks.

- **A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games.**

Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel.

AISTATS, 2020. (Azizian et al., 2020a)

We provide a unified convergence analysis of gradient-based methods for differentiable games across regimes from bilinear to strongly monotone, and prove that extragradient attains an optimal rate in a broad class of extrapolation-based methods.

- **Linear lower bounds and conditioning of differentiable games.**

Adam Ibrahim, Waïss Azizian, Gauthier Gidel, and Ioannis Mitliagkas.

ICML, 2020. (Ibrahim et al., 2020)

We develop linear iteration-complexity lower bounds for differentiable games and introduce a condition number for games, showing that linear convergence rates can arise even without strong convexity or concavity.

- **Accelerating smooth games by manipulating spectral shapes.**

Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel.

AISTATS, 2020. (Azizian et al., 2020b)

We characterize acceleration in smooth games through matrix-iteration theory by introducing spectral shapes, i.e., sets of Jacobian eigenvalues under gradient dynamics, and show that gradient-based methods act as spectral-shape transformations, yielding optimal rates across a continuum of game regimes.

Part I

ON THE LONG-RUN BEHAVIOR OF SGD

OVERVIEW

2.1	Introduction	11
2.2	Overview of our results	12
2.2.1	Our approach: large deviations and random perturbations of dynamical systems	12
2.2.2	Informal setup	12
2.2.3	Asymptotic behavior of SGD	13
2.2.4	Global convergence of SGD	15
2.3	Related Work	16
2.3.1	SGD with a vanishing step-size	16
2.3.2	Diffusion approximation of SGD	16
2.3.3	Stochastic Gradient Langevin Dynamics	17
2.3.4	SGD as a discrete-time Markov chain	18
2.3.5	Exit times for SGD	18
2.3.6	Overparameterized neural networks	19
2.3.7	Large Deviations Principle for stochastic algorithms	20

2.1 INTRODUCTION

Even though SGD has been around for more than 70 years, it is still the method of choice for training a wide array of machine learning architectures – from large language models to reinforcement learning and recommender systems. This phenomenal success is largely owed to the method’s simplicity: given a smooth function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and the associated optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) \quad (\text{Opt})$$

the SGD algorithm is given by the simple update rule

$$x_{n+1} = x_n - \eta \hat{g}_n \quad (\text{SGD})$$

where $\eta > 0$ is the method’s step-size and $\hat{g}_n$ is a stochastic gradient of f at x_n .

By virtue of its wide applicability, (SGD) and its variants have been studied extensively in the literature, for both convex and non-convex objectives. Of these two settings, the non-convex case is the most relevant setting for machine learning, and in particular deep learning, where neural networks lead to highly non-convex optimization problems.

In view of this, our main objective in this part is to shed light on the *long-term* behavior of (SGD) in the non-convex setting. In this part, we seek to shed light on the following central question:

What is the long-term behavior of (SGD) on non-convex objectives?

In particular, we will focus on the following open questions:

- (a) *Where are the iterates of (SGD) most likely to be found on the long run?*
- (b) *How much time does it take for (SGD) to get a region of interest, such as the neighborhood of the global minimum?*

In the next section § 2.2, we present an overview of our main results while in § 2.3 we present a review of the state of the art on related topics.

2.2 OVERVIEW OF OUR RESULTS

2.2.1 Our approach: large deviations and random perturbations of dynamical systems

The linchpin of our approach is the theory of large deviations and random perturbations of dynamical systems of Freidlin and Wentzell (1998b) and Kifer (1988a) for Markov processes, which characterizes the long-run behavior of stochastic processes—invariant measure, hitting times of relevant sets—in the small-noise regime. However, this theory was originally developed for continuous-time diffusions on closed manifolds, whereas (SGD) presents three key challenges:

- It operates in discrete time;
- Its domain is unbounded;
- its noise is not necessarily Gaussian, and may be discrete.

This introduces a series of major technical challenges and obstacles, which we describe in detail in the next chapters. Specifically, the remainder of this section presents an informal version of our main results. The full setting and an outline of the main elements of the proof are provided in Chapter 3 (with the full proofs in Appendix A and Appendix B). In Chapter 4, we provide a formal statement of our main results on the invariant measures of (SGD) along with consequences and extensions, while Chapter 5 is devoted to hitting times.

2.2.2 Informal setup

Regarding the gradient input to (SGD), we will assume throughout that the optimizer has access to a *stochastic first-order oracle* (SFO), that is, a black-box mechanism returning a stochastic estimate of the gradient of f at the point of interest. Formally, when queried at $x \in \mathbb{R}^d$, an SFO returns a random vector of the form

$$G(x; \omega) = \nabla f(x) + Z(x; \omega) \tag{SFO}$$

where

- (a) ω is a random seed drawn from some (complete) probability measure $\mathbb{P}$ supported on a compact subspace of $\mathbb{R}^m$ for some $m \geq 1$.
- (b) $Z(x; \omega)$ is an umbrella error term capturing all sources of noise and randomness in the oracle.

This oracle model is sufficiently flexible to account for all established versions of (SGD) in the literature, including minibatch SGD (where ω represents the sampled minibatch), and noisy gradient descent (where the optimizer may artificially inject noise in the process to enhance convergence).

The precise assumptions for f and (SFO) are detailed later. For now, the mental picture that the reader should have in mind is that of a multimodal non-convex function whose set of critical points

$$\text{crit } f := \{x \in \mathbb{R}^d : \nabla f(x) = 0\} \quad (2.1)$$

consists of a finite number of (smoothly) connected components $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, see Figure 6.

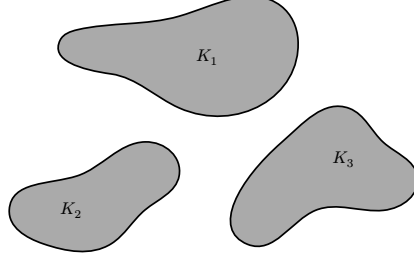


Figure 6: The critical set $\text{crit } f$ of f consists of a finite number of (smoothly) connected components $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$.

2.2.3 Asymptotic behavior of SGD

Let us first focus on the first question raised in the beginning of this part, that is, the long-term behavior of (SGD):

Where are the iterates of (SGD) most likely to be found on the long run?

To answer this question, we first precise what we mean by “long-term” behavior. Since the step-size η is constant, (SGD) is a time-homogeneous Markov process, so the long-run distribution of the iterates of (SGD) can be described by the following notions of stationarity:

Definition 2.1 (Invariant measures). A probability measure μ_∞^η on $\mathbb{R}^d$ is an *invariant measure* of (SGD) if, for any $n \geq 0$, $x_n \sim \mu_\infty^\eta$ implies $x_{n+1} \sim \mu_\infty^\eta$.

Why are we interested in invariant measures? Because they capture the average long-term behavior of (SGD). To make this intuition formal, we consider the occupation measures of (SGD), defined as

Definition 2.2 (Occupation measures). The occupations measures μ_n of (SGD), for $n \geq 0$, are the probability measures on $\mathbb{R}^d$ that, for every Borel $\mathcal{B} \subseteq \mathbb{R}^d$, are given by

$$\mu_n(\mathcal{B}) = \mathbb{E} \left[\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}\{x_k \in \mathcal{B}\} \right] \quad (2.2)$$

The occupation measures μ_n capture the average behavior of (SGD) up to time n . Under mild conditions (see § 4.2.3 or § 4.2.4), the occupation measures μ_n converge in the weak- $\star$ topology¹ to the invariant measures of (SGD) as $n \rightarrow \infty$: informally, for any set of interest $\mathcal{B} \subset \mathbb{R}^d$, for $n \rightarrow \infty$,

$$\mu_n(\mathcal{B}) \approx \mu_\infty^\eta(\mathcal{B}). \quad (2.3)$$

¹ Weak convergence means here that $\lim_{n \rightarrow \infty} \int \varphi d\mu_n = \int \varphi d\mu_\infty^\eta$ for every bounded continuous function $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$. This means that $\mu_n(\mathcal{B})$ will only be asymptotically upper-bounded or lower-bounded by $\mu_\infty^\eta(\mathcal{B})$ depending on whether $\mathcal{B}$ is closed or open, but we dismiss these technicalities in the main text.

Invariant measures thus capture the average long-term behavior of (SGD). Our main result on the long-term behavior of SGD is thus stated in terms of the invariant measures of (SGD).

In view of Definition 2.1 our main results for the long-run distribution of (SGD) can be summarized in the four theorems below, which we state in an informal manner, without the precise definitions of the quantities involved, which we will introduce in the following chapters (for the full version, see § 4.1).

The first theorem establishes that the invariant measures of (SGD) are supported on the critical set $\text{crit } f$ of f .

Theorem 2.1 (Concentration on critical points). *Suppose that μ_∞^η is invariant under (SGD). Then, for small η , we have:*

$$\mu_\infty^\eta(\text{crit } f) \geq 1 - e^{-\frac{c}{\eta}} \quad \text{for some } c > 0 \quad (2.4)$$

This result guarantees that the invariant measures of (SGD) are supported on $\text{crit } f$ with exponentially high probability so we can focus on the behavior of (SGD) near $\text{crit } f$ to understand the long-term behavior of (SGD).

Theorem 2.2 (Boltzmann–Gibbs distribution). *Suppose that μ_∞^η is invariant under (SGD). Then, each critical component C_i of f admits an energy level E_i , $i = 1, \dots, N_{\text{crit}}$, such that, for small η , we have*

$$\mu_\infty^\eta(C_i) \propto \exp\left(-\frac{E_i}{\eta}\right). \quad (2.5)$$

The precise definition of the energy E_i of each component C_i is central in the above theorem, but it requires a fair amount of technical machinery, so we only provide a sketch here and defer the details to § 4.1. At a high level, the main elements of this construction are as follows:

The energy E_i of each component C_i thus determines the relative probabilities of (SGD) to be found near each component of $\text{crit } f$ on the long run. To give some intuition, let us sketch where this energy comes from; the precise definition is in . To each pair of components C_i, C_j of $\text{crit } f$, one can associate a *transition cost* $B_{i,j}$ measuring how hard it is for (SGD) to move from C_i to C_j : it is the minimal cost over all paths joining the two components, where the cost of a path depends on both f and the statistics of (SFO). The energy E_i is then obtained by isolating the minimum-weight spanning tree in the graph whose vertices are the components of $\text{crit } f$ and whose edges are weighted by the $B_{i,j}$. In particular, E_i is a *non-local* quantity: it reflects not only the local geometry of f around C_i , but the way C_i is embedded in the larger landscape of critical components.

In particular, the critical set of f includes several types of critical components: local minima, local maxima and saddles. We will make the distinction between minimizing and *non-minimizing* components: a component C_i of $\text{crit } f$ is called (locally) minimizing if $C_i = \arg \min_{x \in \mathcal{U}} f(x)$ for some neighborhood $\mathcal{U}$ of C_i ; otherwise, we say that C_i is non-minimizing.

Theorem 2.3 (Saddle-point avoidance). *Suppose that μ_∞^η is invariant under (SGD), and let C_i be a non-minimizing critical component of f . Then there exists a minimizing component C_j such that, for small η , we have*

$$\frac{\mu_\infty^\eta(C_i)}{\mu_\infty^\eta(C_j)} \leq e^{-\frac{c}{\eta}} \quad \text{for some } c > 0 \quad (2.6)$$

This theorem establishes that non-minimizing components of $\text{crit } f$ have higher energy than some minimizing component, and thus they are exponentially less likely to be found on the long run than some minimizing component.

Going further, combining [Theorem 2.2](#) with [Theorem 2.1](#), we can show that the invariant measures of (SGD) are supported on one group of components of $\text{crit } f$: the *ground state* of the system, which is the group of components of $\text{crit } f$ with minimal energy:

$$\mathcal{C}_0 := \bigcup_{i \in \arg \min_j E_j} \mathcal{C}_i \quad (2.7)$$

Theorem 2.4 (Ground state concentration). *Suppose that μ_∞^η is invariant under (SGD), and let $\mathcal{C}_0$ be the ground state of f . Then, for small η , we have*

$$\mu_\infty^\eta(\mathcal{C}_0) \geq 1 - e^{-\frac{c}{\eta}} \quad \text{for some } c > 0 \quad (2.8)$$

In words, [Theorems 2.2–2.4](#) provide the following quantification of the limiting distribution of x_n : in the long run

- (a) the critical region of f is visited exponentially more often than any non-critical region of f ([Theorem 2.1](#));
- (b) in particular, the iterates of (SGD) are exponentially concentrated around the problem's ground state ([Theorem 2.4](#));
- (c) among the mass that remains, every component of f gets a fraction that is exponentially proportional to its energy ([Theorem 2.2](#)); and, finally
- (d) every non-minimizing component is “dominated” by a minimizing component that is visited exponentially more often ([Theorem 2.3](#)).

Remark 1. Importantly, the problem's energy landscape is shaped by f , but not f alone: the statistics of (SFO) play an equally important role, so we may have $\mathcal{C}_0 \neq \arg \min f$. 🐼

2.2.4 Global convergence of SGD

In this section, we focus on the second question raised in the beginning of this part, that is, the global convergence of (SGD):

How much time does it take for (SGD) to reach a neighborhood of the global minimum?

To answer this question, assume in the sequel that (SGD) is initialized at some fixed $x_0 \leftarrow x$. Then, given a target tolerance $\delta > 0$, we will estimate the time required for (SGD) to get within δ of $\arg \min f$, that is, the hitting time

$$\tau = \inf\{n \in \mathbb{N} : \text{dist}(x_n, \arg \min f) \leq \delta\} \quad (2.9)$$

Our main result provides matching upper and lower bounds on the expected hitting time $\mathbb{E}_x[\tau]$ of (SGD) started at x to $\arg \min f$.

Theorem 2.5. *Suppose that (SGD) is initialized at $x \in \mathbb{R}^d$. Then, given a tolerance level $\varepsilon > 0$, and small enough $\delta, \eta > 0$, we have*

$$\exp\left(\frac{E(x) - \varepsilon}{\eta}\right) \leq \mathbb{E}_x[\tau] \leq \exp\left(\frac{E(x) + \varepsilon}{\eta}\right).$$

In the above, $E(x)$ is an energy value, not unlike the energy levels appearing in [Theorem 2.2](#) (and constructed in a similar manner). In this regard, it measures

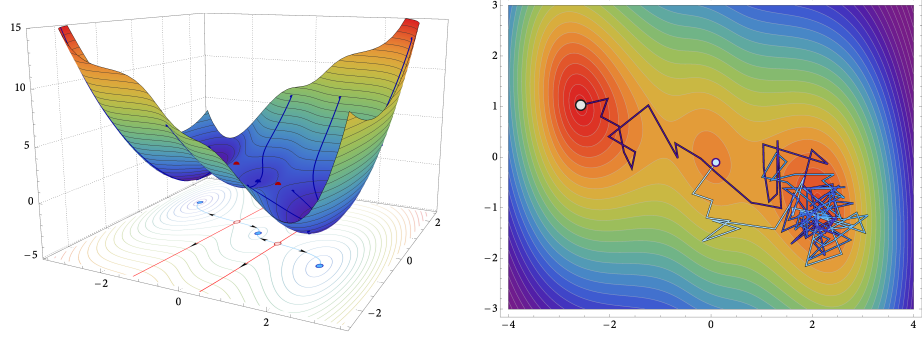


Figure 7: Two views of the three-hump camel benchmark: the loss landscape (left) and a representative SGD trajectory toward the global minimizer (right).

how hard it is for SGD to reach the global minimum $\arg \min f$ starting from x . It is a geometric measure of the problem’s hardness and we will see later that it captures both the difficulty of the loss landscape and the noise profile of (SFO) in a similar way as the energies E_i of Theorem 2.2.

2.3 RELATED WORK

2.3.1 SGD with a vanishing step-size

The long-run behavior of (SGD) is markedly different when the method is run with a vanishing step-size $\eta_n > 0$ with $\lim_{n \rightarrow \infty} \eta_n = 0$. This was the original version of (SGD) as proposed by Robbins and Monro (1951) – and, in a slightly modified form by Kiefer and Wolfowitz (1952) – and, in contrast to the constant step-size case, the vanishing step-size algorithm converges with probability 1. The first almost sure convergence result obtained by Ljung (1977) under the assumption that the iterates of (SGD) remain bounded. This boundedness assumption was dropped by Benaïm (1999) and Bertsekas and Tsitsiklis (2000) who showed that (SGD) with a vanishing step-size converges to a component of $\text{crit } f$ as long as η_n satisfies the Robbins–Monro summability conditions $\sum_n \eta_n = \infty$ and $\sum_n \eta_n^2 < \infty$; for a series of related results under different assumptions, cf. Zhou et al. (2020), Mertikopoulos et al. (2020, 2024) and references therein. In all these cases, ergodicity of the process is lost (because of the vanishing step-size), so the limiting distribution of (SGD) depends crucially on its initialization and other non-tail events.

2.3.2 Diffusion approximation of SGD

A classical thread in the literature – and the foundation of Langevin sampling methods – revolves around gradient flows perturbed by Brownian noise in continuous time; in this case, it is possible to get an explicit expression for the invariant measure of the resulting stochastic differential equation (SDE) Jordan et al. (1998), or estimate it in more difficult, constrained cases Benaïm (1999), Borkar (2008), Zhou et al. (2017, 2020), Mertikopoulos and Staudigl (2018a,b). Owing to our understanding of gradient-based SDEs, an important thread in the literature concerns the so-called *SDE approximation* of (SGD), as pioneered by Li et al. (2017, 2019). Applications of this SDE approximation include the study of the dynamics of (SGD) close to manifolds of minimizers (Blanc et al., 2020, Li et al., 2021) or Yang et al. (2020) which quantifies the global

convergence of the diffusion approximation of (SGD) when the function has no spurious local minima. Wojtowytsch (2023b) study the invariant measure of the diffusion approximation: if the set of global minimizers form a manifold on which the noise vanishes, they show that the invariant measure of the diffusion concentrates on this manifold and moreover provide a description of the limiting measure on this manifold. When the objective function is scale-invariant (Li et al., 2020b), we can obtain further results: Wang and Wang (2022) describes the convergence of the SDE approximation to its asymptotic regime, while Li et al. (2021, 2022b) also quantify the convergence of (SGD) initialized close to minimizers. Li et al. (2024a) refine this approximation further by developing a Hessian-aware stochastic modified equation that more accurately captures the local behaviour of SGD. Li et al. (2024b) study the implicit bias of the SDE approximation of SGD through a large deviation analysis, showing how, on the manifold of global minimizers, the invariant measure of the diffusion depends on both the curvature of the objective and the geometry of this manifold.

Another line of work leverages dynamic mean-field theory (DMFT) to study the behavior of the diffusion approximation of (SGD) (Mignacco et al., 2020, Mignacco and Urbani, 2022, Veiga et al., 2024). The DMFT, or “path-integral” approach, comes from statistical physics and bears a close resemblance to the Freidlin-Wentzell theory of large deviations for SDEs. However, this methodology, as well as Mignacco et al. (2020), Mignacco and Urbani (2022), Veiga et al. (2024), is restricted to the continuous-time diffusion and remains at heuristic level. More recently, Martin et al. (2026) provide a high-dimensional analysis of the gradient flow and the Langevin process for extensive-width quadratic neural networks, obtaining exact characterizations of the training dynamics in this setting. All these results are either local or concern the asymptotic behaviour of the continuous-time approximation of (SGD). They do not provide information of the global convergence time of the actual discrete-time (SGD), since the approximation guarantees fail to hold on large enough time intervals.

2.3.3 Stochastic Gradient Langevin Dynamics

Raginsky et al. (2017) analyzed the law of the iterates of stochastic gradient Langevin dynamics (SGLD), a discrete-time sampling algorithm designed to draw samples from a Gibbs measure $\propto e^{-\beta f}$ via the update rule:

$$x_{n+1} = x_n - \eta \hat{g}_n + \sqrt{2\eta\beta^{-1}} \xi_n, \quad (2.10)$$

where $\hat{g}_n$ is a stochastic gradient of f at x_n and ξ_n is standard Gaussian noise. They establish that, for sufficiently small step size η , the discrete dynamics (2.10) track the corresponding diffusion process closely in Wasserstein distance over finite time horizons. Exploiting the geometric convergence of this diffusion to its invariant measure—the Gibbs measure $\propto e^{-\beta f}$ —and the fact that this measure satisfies a logarithmic Sobolev inequality (LSI), they obtain convergence guarantees for SGLD.

Unfortunately, the resulting convergence rate has exponentially poor dependence on β , which renders this approach unsuitable for analyzing SGD. Approximating (SGD) by (2.10) requires setting $\beta \sim 1/\eta$, so as $\eta \rightarrow 0$ to validate the diffusion approximation, β diverges. The LSI constant of the Gibbs measure $\propto e^{-\beta f}$ degrades exponentially in β , so the convergence bound blows up at the

very regime where the approximation becomes accurate, rendering the overall guarantee vacuous.

Follow-ups include [Erdogdu et al. \(2018\)](#), who propose to leverage such algorithms for global optimization of non-convex functions, [Li and Wang \(2022\)](#), who show that the error in Kullback–Leibler (KL) divergence between the discrete and continuous processes is bounded uniformly in time when the objective is strongly convex. Another line of works leverages the theory of [Eberle and Majka \(2019\)](#), [Eberle et al. \(2017, 2019\)](#), [Eberle \(2016, 2011\)](#) to obtain non-asymptotic bounds on the convergence of SGLD to its invariant measure in Wasserstein distance, without invoking the LSI ([Majka et al., 2020](#), [Lamperski, 2021](#), [Cheng et al., 2020](#)). However, all these approaches still require setting $\beta \sim 1/\eta$ to approximate SGD, so as $\eta \rightarrow 0$ the temperature diverges and the convergence bounds blow up for the same reason.

A related line of work studies how the injecting some noise into SGD can shape the training dynamics and generalization. [Wu et al. \(2020\)](#) and [Zheng et al. \(2021\)](#) show that perturbation of model weights explicitly encourages convergence to flatter minima, improving generalization. [Lucchi et al. \(2022\)](#) study the theoretical properties of noise correlation in stochastic optimization; [Orvieto et al. \(2022\)](#) show that anticorrelated noise injection biases the dynamics toward wider minima and improves generalization, and [Orvieto et al. \(2023\)](#) show that it induces explicit regularization in overparameterized models.

2.3.4 SGD as a discrete-time Markov chain

In a relatively recent thread in the literature, [Dieuleveut et al. \(2020\)](#) and [Lu et al. \(2021\)](#) undertook a study of (SGD) as a discrete-time Markov chain. This allowed them to derive conditions under which (SGD) is (geometrically) ergodic and, in this way, to quantify the distance to the minimizer under global growth conditions. Building further on this perspective, [Gurbuzbalaban et al. \(2021\)](#), [Hodgkinson and Mahoney \(2021\)](#) and [Pavasovic et al. \(2023\)](#) showed that, under general conditions, the asymptotic distribution of the iterates of (SGD) is heavy-tailed; As such, these results concern the probability of observing the iterates of (SGD) at very large distances from the origin. This is in contrast with our work, which focuses on the time it takes (SGD) to reach a global minimizer. These two types of results are orthogonal and complementary to each other.

2.3.5 Exit times for SGD

Regarding the second question, we will study how long it takes for (SGD) to reach a global minimizer. There is vast literature that instead seeks to understand how long it takes for (SGD) to exit a local minimum. For this, these works study either the diffusion approximation ([Xie et al., 2020](#), [Hu et al., 2019](#), [Mori et al., 2022](#), [Jastrzębski et al., 2017](#), [Gesù et al., 2019](#), [Baudel et al., 2023](#)) or on heavy-tail versions of this diffusion ([Nguyen et al., 2019b](#), [Wang et al., 2022b](#)). Interestingly, some of these works use elements of the continuous-time Freidlin–Wentzell theory ([Freidlin and Wentzell, 1998b](#)), which is also the point of departure of our work.

Note that studying exit times is not enough to characterize the convergence time of (SGD) to a global minimizer. In fact, a simple way to see this is given by the example in § 5.1. Indeed, in this case, the escape time of p_3 scales exponentially in $f(p_2) - f(p_3)$ (the depth of p_3) while the escape time of p_5 would scale exponentially in $f(p_4) - f(p_5)$. However, the global convergence time of

(SGD) as determined by (5.3) is greater than either of these two exponentials: indeed, our result takes into account that, when in the vicinity of p_3 , (SGD) could transition to p_5 instead of p_1 an arbitrary number of times. Hence, characterizing escape times is not enough to describe the global convergence time of (SGD) in general.

2.3.6 Overparameterized neural networks

The pioneering work of (Allen-Zhu et al., 2019, Du et al., 2019, Zou et al., 2020a) show some convergence results for gradient methods on neural networks: overparameterization and Gaussian initialization enable convergence to a global minimum. Among the many subsequent works, let us mention (Azizan et al., 2022, Weinan et al., 2019, Zhang et al., 2022a, Zou et al., 2020b, Zou and Gu, 2019). In particular, with N denoting the number of training datapoints, this type of results require either a $O(N^2)$ -width with Gaussian initialization (Oymak and Soltanolkotabi, 2020, Nguyen and Mondelli, 2020) or a $O(N)$ -width under additional structure: infinite depth (Marion et al., 2024) or specific initialization (Nguyen and Mondelli, 2020, Liu et al., 2023). Our work addresses a different question as it provides a convergence analysis on general non-convex functions, regardless of the structure of the objective or the initialization.

A fruitful line of works studies the geometric properties of the loss landscapes of neural networks, and in particular whether they contain spurious local minima (Kawaguchi, 2016). Nguyen and Hein (2017), Nguyen et al. (2019a) show that there are no spurious local minima, or “bad valley”, if the architecture possesses an hidden layer with width at least N . There are many refinements: e.g., leveraging the structure of the data (Venturi et al., 2019), considering regularization (Wang et al., 2022c), convolutional neural networks (Nguyen and Hein, 2018) or other variations of the architecture (Wang et al., 2022c, Pollaci, 2023, Arjevani and Field, 2022, Petzka and Sminchisescu, 2021). Moreover, Li et al. (2022a) shows that this requirement is tight for one-hidden layer neural networks, in some settings. When the sizes of the hidden layers are smaller than N , the loss landscape does generally have spurious local minima (Christof and Kowalczyk, 2023). Interestingly, in this case, the depth of these spurious minima can be explicitly bounded (Nguyen et al., 2021) and made arbitrarily small with hidden layers of size only $\sqrt{N}$.

Another fundamental property of overparameterized neural networks is that they can perfectly fit the training data and therefore, the noise of (SGD) vanishes at the global minimum. In this setting, under sufficient noise assumption everywhere else, Wojtowytsch (2023a) shows that (SGD) reaches a neighborhood of a global minimizer almost surely, gets trapped in this neighborhood and then converges to the global minimizer with a provided asymptotic rate. Our work can be seen as a precise estimation of the time to reach the neighborhood of the global minimizer.

Also motivated by the interpolation phenomenon, Islamov et al. (2024) introduces the so-called $\alpha - \beta$ condition and shows a global convergence of (SGD) under this condition. Though this condition is much weaker than convexity, it only ensures that the function value of the iterates becomes less than the max of the values over all critical points. In contrast, our work characterizes the time to convergence to global minimizer for any non-convex landscape.

A complementary perspective comes from mean-field theory. Chizat and Bach (2018) study the global convergence of gradient descent for overparameterized

models via optimal transport. [Barboni et al. \(2022\)](#) extend this to residual networks via a linear parametrization, connecting finite-width ResNets to infinite-width limits through a reproducing kernel Hilbert space framework. [Barboni et al. \(2025\)](#) further develop a mean-field framework for simultaneously infinitely deep and wide ResNets using conditional optimal transport. [Petit et al. \(2026\)](#) extend this framework to wide shallow models with bounded nonlinearities — including attention layers and sigmoid networks — and show that non-global minimizers are unstable under the limiting gradient-flow dynamics. [Chizat et al. \(2019\)](#) identify a *lazy training* phenomenon in which wide networks remain close to their initialization throughout training, reducing to a kernel method. [Chizat and Bach \(2020\)](#) characterize the implicit bias of gradient descent for wide two-layer networks trained with the logistic loss. More recently, [Dana et al. \(2025\)](#) establish convergence guarantees for shallow networks when the data is weakly interacting. Finally, [Marcotte et al. \(2023\)](#) develop a general framework for conservation laws in gradient-flow training: these are quantities determined by the model parametrization that remain invariant along all training trajectories, independently of the data and loss. [Marcotte et al. \(2025\)](#) extend this analysis to residual networks and Transformers.

2.3.7 Large Deviations Principle for stochastic algorithms

Large deviation is a large field of research which studies the asymptotic probabilities of rare events, see [\(Dembo and Zeitouni, 2010\)](#) for a textbook. Freidlin and Wentzell, through their seminal works on this subject — see [Freidlin and Wentzell \(1998a\)](#) and references therein —, have developed a large deviation approach to diffusion processes in continuous time on closed manifolds and, more generally, Markov processes. It has seen numerous extensions, in particular some in discrete time ([Kifer, 1988b](#), [Gulinsky and Veretennikov, 1993](#), [de Acosta, 2022](#), [Ganguly, 2018](#), [Dupuis, 1988](#)) with applications to simulated annealing ([Catoni, 1991, 1992](#), [Cerf, 1996](#), [Trouné, 1996](#)), stochastic approximation ([Dupuis and Kushner, 1985, 1987](#), [Dupuis, 1988](#), [Hult et al., 2023](#)) and high-probability bounds for strongly convex gradient descent ([Bajovic et al., 2022](#)). The large deviation principle which we is derived from [Freidlin and Wentzell \(1998a\)](#) and, though it could have been obtained from other works ([Dupuis, 1988](#), [Ganguly, 2018](#)) with similar assumptions, we chose to follow their approach as it is convenient, general and the most well-known.

THE LARGE DEVIATIONS VIEWPOINT

3.1	Large deviations, long-run behavior, and gradient flow	22
3.1.1	The Freidlin–Wentzell–Kifer setting	22
3.1.2	Large deviations and the action functional	23
3.1.3	The approach of Freidlin–Wentzell and Kifer	23
3.1.4	Our extension to SGD	24
3.2	Setup	25
3.2.1	Setup and assumptions	25
3.2.2	Examples: regularized empirical risk minimization	28
3.3	Large deviations for SGD	29
3.3.1	Intuition	29
3.3.2	Hamiltonian, Lagrangian and action functional	30
3.3.3	Large deviation principle for “continuous-time SGD trajectories”	32
3.3.4	Large deviation principle for accelerated SGD	33
3.4	The induced chain	35
3.5	Reduction to a finite state-space	36
3.5.1	Equivalence classes of points in $\text{crit } f$	37
3.5.2	Regularity of the Lagrangian and the rate function near critical points	38
3.5.3	Equivalence classes coincide with connected components of $\text{crit } f$	39
3.5.4	Putting everything together	39
3.6	Transition times	41
3.6.1	Lyapunov potential	42
3.6.2	Analysis near the critical region	43
3.6.3	Estimating the transition times	44
3.7	Transition probabilities	45
3.7.1	Transition costs	45
3.7.2	Transition probabilities	46
3.8	Summary	48

In this chapter, we present a detailed sketch of the proof of our main results on SGD presented in the previous chapter. Though the proofs are presented in the appendices of [Azizian et al. \(2024c, 2025a\)](#), the presentation in this chapter is entirely new, with expanded details, explanations, and perspectives. The goal of this chapter is thus to provide the intuition of what makes the approach work.

In § 3.1, we give an informal overview of the proof strategy, and explain how it relates to the classical theory of [Freidlin and Wentzell \(1998b\)](#) and [Kifer \(1988a\)](#). In § 3.2, we introduce the setup and discuss the assumptions. In § 3.3, we present the large deviation principle for SGD, which is the linchpin of

our approach. Then, we flesh out how to transfer the approach of Freidlin and Wentzell (1998b) and Kifer (1988a) to SGD: in § 3.4, we define the induced chain for SGD and show that it can be seen as a finite-state Markov chain by studying the structure of the set of critical points in § 3.5; in § 3.6, we characterize the transition times of SGD between components of critical points while, finally, in § 3.7, we obtain precise estimates on the transition probabilities of the induced chain. These elements are the key ingredients for the proofs of the main results on the invariant measure and the global convergence time of SGD, which are presented in the next chapters Chapter 4 and Chapter 5.

3.1 LARGE DEVIATIONS, LONG-RUN BEHAVIOR, AND GRADIENT FLOW

Before presenting our setup and the technical details of the proof, we sketch the overall strategy and explain how it relates to the classical theory of Freidlin and Wentzell (1998b) and Kifer (1988a). The goal of this section is to give the reader a clear mental picture of what the proof achieves, and why each of the three sections that follow is necessary.

3.1.1 The Freidlin–Wentzell–Kifer setting

The starting point of Freidlin and Wentzell (1998b) and Kifer (1988a) is the study of a randomly perturbed dynamical system on a *closed* (compact without boundary) manifold. The prototypical example is the stochastic differential equation¹

$$dX_t = -\nabla f(X_t) dt + \sqrt{\varepsilon} \sigma dW_t, \quad (\text{SDE})$$

where W_t is a standard Brownian motion, $\sigma > 0$ is a noise coefficient, and $\varepsilon > 0$ is a small parameter playing the role of the step size. This is the continuous-time counterpart of (SGD): the drift $-\nabla f$ is the deterministic part, and $\sqrt{\varepsilon} \sigma dW_t$ is the random perturbation. Our analysis of (SGD) is directly inspired by the approach developed by Freidlin and Wentzell (1998b) and Kifer (1988a) for (SDE): by leveraging large deviations theory, these works aim at characterizing the long-run behavior of (SDE) when the parameter ε goes to 0. Naturally, the behavior of the perturbed process (SDE) when ε is small is closely related to the behavior of the unperturbed dynamics, which is given by the gradient flow of f .

GRADIENT FLOW. A central role in the analysis is played by the *gradient flow* of f , i.e., the deterministic dynamics obtained from (SDE) by removing the noise.

Definition 3.1 (Gradient flow and flow map). A gradient flow of f is a curve $\gamma: \mathbb{R}_+ \rightarrow \mathbb{R}^d$ satisfying the ordinary differential equation

$$\dot{\gamma}(t) = -\nabla f(\gamma(t)) \quad (3.1)$$

Define, for $x \in \mathbb{R}^d$, the *flow map* $\Theta_t(x)$ as the solution of the gradient flow starting at x at time 0, i.e.,

$$\dot{\Theta}_t(x) = -\nabla f(\Theta_t(x)) \quad \text{with} \quad \Theta_0(x) = x. \quad (3.2)$$

¹ Formally, we should consider a SDE on a manifold, but we write it in $\mathbb{R}^d$ for simplicity since it is not our focus here.

The gradient flow is the “backbone” of the dynamics of (SDE) when ε is small: the trajectories of (SDE) are likely to be close to those of the gradient flow. However, due to the noise, the process can deviate from the gradient flow and even escape from its attractors, which is what drives the long-run behavior of (SDE). These deviations are captured by the large deviation theory, which quantifies how likely it is for (SDE) to follow a given trajectory γ that may deviate from the gradient flow.

3.1.2 Large deviations and the action functional

Though, on average, the trajectories of (SDE) follow the gradient flow, they can deviate from it due to the noise. The key tool for studying these deviations and is a *large deviation principle* for sample paths.

Given a curve $\gamma: [0, T] \rightarrow \mathbb{R}^d$, the large deviation principle (LDP) quantifies how likely it is for the process $(X_t)_{t \leq T}$ to follow γ when ε is small. In a very informal fashion, it reads:

$$\mathbb{P}((X_t)_{t \leq T} \approx \gamma) \approx \exp\left(-\frac{\mathcal{S}_T[\gamma]}{\varepsilon}\right), \quad (3.3)$$

where $\mathcal{S}_T[\gamma]$ is the *action functional* defined by:

$$\mathcal{S}_T[\gamma] = \int_0^T \frac{\|\dot{\gamma}(t) + \nabla f(\gamma(t))\|^2}{2\sigma^2} dt. \quad (3.4)$$

Thus (3.3) fully characterizes the likelihood of any trajectory γ for the perturbed process (SDE) when ε is small: this probability is exponential in $1/\varepsilon$, and the exponent is given by the action functional $\mathcal{S}_T[\gamma]$. The action functional measures how much γ deviates from being a solution of the gradient flow. It satisfies:

$$\mathcal{S}_T[\gamma] \geq 0 \quad \text{and} \quad \mathcal{S}_T[\gamma] = 0 \iff \dot{\gamma}(t) = -\nabla f(\gamma(t)) \text{ for a.e. } t \in [0, T], \quad (3.5)$$

i.e., it vanishes exactly on solutions of the gradient flow. Therefore, the farther γ is from a gradient flow trajectory, the higher $\mathcal{S}_T[\gamma]$ is, and the less likely the perturbed process is to follow γ . Conversely, trajectories close to the gradient flow are highly probable when ε is small. But, crucially, in both cases, (3.3) gives a precise quantification of the likelihood of γ for the perturbed process (SDE) when ε is small, which is the key to understanding the long-run behavior of (SDE).

3.1.3 The approach of Freidlin–Wentzell and Kifer

Armed with the LDP (3.3), Freidlin and Wentzell (1998b) and Kifer (1988a) analyze the long-run behavior of (SDE) by introducing a certain *induced chain* obtained by observing (SDE) only at the times it visits neighborhoods of the critical components of f . Their analysis relies on three key steps:

STEP 1: ACCURATE MODELLING BY THE INDUCED CHAIN. One of the main challenges is first to show that this induced chain, obtained by observing (SDE) only at the times it visits neighborhoods of the critical components, actually represents the long-run behavior of the full dynamics. For this, one needs to show that the perturbed process spends most of its time near the critical components of f : this is achieved by a precise control of the distribution of the

time spent away from the critical components, in other words, of the transition times between critical components.

STEP 2: REDUCTION TO A FINITE-STATE SPACE. Another key ingredient is to show that it is possible to abstract away what happens inside the critical components themselves. Instead of treating the process as evolving on the manifold, one can show that it behaves, to leading order in ε , as a Markov chain on the critical components of f . More precisely, the behavior of the process started at x is roughly the same as that started at another point x' if x and x' are in the same critical component C_i . This relies on the regularity of the Lagrangian and action functional near the critical components, which is significantly more tractable in the compact manifold setting with Brownian noise.

STEP 3: TRANSITION PROBABILITIES. The third ingredient is obtaining precise estimates on the transition probabilities of the induced chain: if the process is near some C_i , how likely is it to reach C_j next? By the LDP (3.3), these probabilities are exponential in ε :

$$\mathbb{P}(\text{process goes from } C_i \text{ to } C_j) \sim \exp\left(-\frac{B_{i,j}}{\varepsilon}\right), \quad (3.6)$$

where B_{ij} is the *transition cost* from C_i to C_j , defined as the minimum action to travel from C_i to C_j :

$$B_{ij} = \inf\{S_T[\gamma] : \gamma \text{ goes from } C_i \text{ to } C_j \text{ for some } T > 0\}. \quad (3.7)$$

These transition costs encode all the relevant geometry of f and statistics of the noise, and they drive the long-run distribution and the convergence time.

3.1.4 Our extension to SGD

The linchpin of our approach is to carry out this same three-step program for (SGD), which differs from (SDE) in three important points:

- (i) It is a discrete-time system, while (SDE) is a continuous-time system.
- (ii) It evolves on an unbounded domain ($\mathbb{R}^d$), while (SDE) is defined on a closed manifold (compact manifold without boundary).
- (iii) Its noise is general (non-Gaussian and possibly discrete), while (SDE) is driven by Brownian noise.

All these differences create substantial technical difficulties. The main challenge is that most of the potentials introduced in Freidlin and Wentzell (1998b), Kifer (1988a) become drastically less regular in our context. We remedy this issue by refining the analysis and carefully studying the structure of the attractors of (SGD). This allows us to salvage enough regularity and show that (SGD) spends most of its time near its critical components (this is achieved by developing suitable tail bounds for the time spent away from critical points), and, ultimately, to estimate the transition probabilities of (SGD) between different connected components of critical points.

The rest of this chapter is organized as follows. We first set up notation and assumptions in § 3.2. We then present the LDP for (SGD) in § 3.3: this is the discrete-time, unbounded-domain version of (3.3). Finally, we carry out the three-step program outlined above in § 3.4, § 3.6, and § 3.7. These sections establish the key estimates needed for the proofs of the main results on the

invariant measure and the global convergence time of SGD, which are presented in the next chapters [Chapter 4](#) and [Chapter 5](#).

3.2 SETUP

3.2.1 Setup and assumptions

Throughout the sequel, we will write $\langle \cdot, \cdot \rangle$ for the standard inner product on $\mathbb{R}^d$ and $\|\cdot\|$ for the Euclidean norm. We will also write $\mathbb{B}(x, r)$ (resp. $\bar{\mathbb{B}}(x, r)$) for the open (resp. closed) ball of radius r centered at $x \in \mathbb{R}^d$, and we will respectively denote the open and closed δ -neighborhoods of $A \subset \mathbb{R}^d$ as

$$\mathcal{U}_\delta(A) := \{x \in \mathbb{R}^d : \text{dist}(A, x) < \delta\} \quad (3.8a)$$

$$A_\delta := \text{cl}(\mathcal{U}_\delta(A)) = \{x \in \mathbb{R}^d : \text{dist}(A, x) \leq \delta\}, \quad (3.8b)$$

where dist denotes here the distance of a point to a set and cl denotes the closure operator.

We can now introduce our main setup. We begin with our assumptions for the objective function f of [\(Opt\)](#). The first two conditions are standard in optimization.

Assumption 3.2. The objective function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the following conditions:

- (a) *Coercivity:* $f(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$.
- (b) *Smoothness:* f is C^2 -differentiable and its gradient is β -Lipschitz continuous, that is,

$$\|\nabla f(x') - \nabla f(x)\| \leq \beta \|x' - x\| \quad \text{for all } x, x' \in \mathbb{R}^d. \quad (3.9)$$

- (c) *Critical set regularity:* The critical set

$$\text{crit } f := \{x \in \mathbb{R}^d : \nabla f(x) = 0\} \quad (2.1)$$

of f consists of a finite number of connected components $\mathcal{K}_i, i = 1, \dots, K$ and each of these components is *essentially smoothly connected*: for any two points x, x' in the same component $\mathcal{K}_i$, there exists a continuous, almost everywhere differentiable curve $\gamma: [0, 1] \rightarrow \mathcal{K}_i$ such that $\gamma(0) = x$, $\gamma(1) = x'$ and a partition $0 = t_0 < \dots < t_n < \dots < t_N = 1$ of $[0, 1]$ such that γ is integrable on every closed interval of (t_{n-1}, t_n) for all $n = 1, \dots, N$.

The assumption on the critical set of f is less standard: it is a regularity condition that is required for our analysis to go through. We first require that the critical set of f has a finite number of connected components. Thanks to the coercivity requirement, this is equivalent to requiring that $\text{crit}(f)$ has a finite number of connected components in any compact set. This prevents pathological cases such as f having non-isolated critical points that accumulate at in a compact set (e.g., $f(x) = \sin(1/x)$ for $x \neq 0$ and $f(0) = 0$).

Connectedness gives that each connected component $\mathcal{K}$ of $\text{crit } f$ is path-connected: two points $x, x' \in \mathcal{K}$ can be connected by a continuous curve $\gamma: [0, 1] \rightarrow \mathcal{K}$ such that $\gamma(0) = x$, $\gamma(1) = x'$. We require something slightly stronger: that the curve γ can be chosen to be almost everywhere differentiable and integrable, except for a finite number of points. This allows $\mathcal{K}$ to have

“corners” or “cusps”, but only a finite number of them. This thus prevents pathological cases (such as $\mathcal{K}$ being a space-filling curve or a fractal set). In particular, isolated critical points, smooth manifolds, and finite unions of closed manifolds satisfy this requirement. More generally, [Assumption 3.2\(c\)](#) is satisfied whenever f is analytic, semi-algebraic or definable in an o-minimal structure – in which case $\text{crit } f$ is also definable in an o-minimal structure, so each component can be connected by piecewise smooth paths ([Coste, 1999](#), [Van Den Dries and Miller, 1996](#)).

Moving forward, to align our notation with standard conventions in large deviations theory, it will be more convenient to work with $-Z(x; \omega)$ instead of $Z(x; \omega)$ from [\(SFO\)](#) in our proofs. We thus work with the noise process

$$U(x, \omega) = -Z(x; \omega) \quad (3.10)$$

With this notation, we restate the definition of the sequence of SGD with the explicit form of the SFO: the iterates of [\(SGD\)](#), started at $x \in \mathbb{R}^d$, are defined by the following recursion:

$$\begin{cases} x_0 \in \mathbb{R}^d \\ x_{n+1} = x_n - \eta \nabla f(x_n) + \eta U_n, \quad \text{where } U_n = U(x_n, \omega_n) \end{cases} \quad (3.11)$$

where $(\omega_n)_{n \geq 0}$ is a sequence of random variables in $\mathbb{R}^m$. We will denote by $\mathbb{P}_x$ the law of the sequence $(\omega_n)_{n \geq 0}$ when the initial point is x and by $\mathbb{E}_x$ the expectation with respect to $\mathbb{P}_x$.

Assumption 3.3. The noise term $U: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}^d$, with Ω a compact subset of some finite-dimensional vector-space, satisfies the following properties:

- (a) *Properness:* $\mathbb{E}[U(x, \omega)] = 0$ and $\text{cov}(U(x, \omega)) > 0$ for all $x \in \mathbb{R}^d$.
- (b) *Smooth growth:* $U(x, \omega)$ is C^2 -differentiable and satisfies the growth condition

$$\sup_{x, \omega} \frac{\|U(x, \omega)\|}{1 + \|x\|} < \infty. \quad (3.12)$$

- (c) *Sub-Gaussian tails:* The cumulant-generating function of U is bounded as

$$\log \mathbb{E} \left[e^{\langle p, U(x, \omega) \rangle} \right] \leq \frac{1}{2} \sigma_\infty^2(f(x)) \|p\|^2 \quad \text{for all } p \in \mathbb{R}^d, \quad (3.13)$$

where $\sigma: \mathbb{R} \rightarrow (0, \infty)$ is continuous, bounded away from zero, and grows at infinity as $\sigma_\infty^2(f(x)) = \Theta(\|x\|^s)$ for some $s \in [0, 2]$, i.e.,

$$0 < \liminf_{\|x\| \rightarrow \infty} \frac{\sigma_\infty^2(f(x))}{\|x\|^s} \leq \limsup_{\|x\| \rightarrow \infty} \frac{\sigma_\infty^2(f(x))}{\|x\|^s} < \infty. \quad (3.14)$$

Note that, because f is coercive, the dependence of σ_∞^2 on $f(x)$ allows for noise processes with an unbounded variance as $\|x\| \rightarrow \infty$; the specific functional dependence through $f(x)$ instead of x is a modeling choice which we make because it greatly simplifies the proof and our calculations.

Remark 2 (Link with the ABC condition). [Assumption 3.3\(c\)](#) allows for the variance proxy σ_∞^2 to grow as f grows. The two most important cases are $\sigma_\infty^2(f(x))$ being constant or growing as $f(x)$. The latter case is reminiscent of the *expected smoothness* ([Gower et al., 2019, 2021](#)) or *ABC condition* ([Khaled and](#)

Richtárik, 2023) — see Alacaoglu et al. (2025) for a historical review — which, with our notations, read

$$\mathbb{E}[\|\nabla f(x) - U(x, \omega)\|^2] \leq A + B\|\nabla f(x)\|^2 + C(f(x) - f^*), \quad (3.15)$$

which, using Young's inequality and smoothness of f implies (e.g., Garrigos and Gower (2023, Lem. 2.28)) that

$$\begin{aligned} \mathbb{E}[\|U(x, \omega)\|^2] &\leq 2A + 2(B+1)\|\nabla f(x)\|^2 + 2C(f(x) - f^*) \\ &\leq 2A + (4(B+1) + 2C)(f(x) - f^*). \end{aligned} \quad (3.16)$$

Our condition thus allows such growth of the scale of the noise, albeit with a sub-Gaussian tail requirement instead of a second-moment requirement.

Note however that [Assumption 3.4](#) constrains the growth of σ_∞^2 at infinity: $\sigma_\infty^2(f(x))$ cannot grow faster than $\|\nabla f(x)\|^2$ at infinity. Hence, if $\sigma_\infty^2(f(x))$ grows as $f(x)$, then $\|\nabla f(x)\|^2$ must grow at least as fast as $f(x)$ at infinity, which is a strong requirement on the geometry of f at infinity. $\heartsuit$

We also require a signal-to-noise ratio (SNR) condition: it ensures that, at infinity, the signal dominates the noise. This ensures that the iterates of (SGD) do not diverge to infinity.

Assumption 3.4. The signal-to-noise ratio of G is bounded as

$$\liminf_{\|x\| \rightarrow \infty} \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} > 16 \log 6 \cdot d. \quad (\text{SNR})$$

This assumption will help to ensure that SGD does not spend too much time at infinity, we will see that it is commonly satisfied in standard settings in § 3.2.2.

Remark 3 (Geometric version of [Assumption 3.3](#) and [Assumption 3.4](#)). One can get rid of the d -based requirement in [Assumption 3.4](#) by refining the noise assumption of [Assumption 3.3](#). Indeed, switching from a scalar sub-Gaussian proxy to a matrix-valued (anisotropic) sub-Gaussian proxy, one can replace the d -based requirement by a requirement that depends on the effective rank of the noise covariance matrix at infinity

More precisely, if [Assumption 3.3](#) is replaced by the following matrix-valued version: for every $x \in \mathbb{R}^d$ and every $p \in \mathbb{R}^d$,

$$\log \mathbb{E}[\exp(\langle p, U(x, \omega) \rangle)] \leq \frac{1}{2} \langle p, \Sigma_\infty(x) p \rangle.$$

and

$$\text{tr}(\Sigma_\infty(x)) + 2 \log 6 \left(\sqrt{\text{tr}(\Sigma_\infty(x)^2)} + \|\Sigma_\infty(x)\|_{\text{op}} \right) \leq \sigma_\infty^2(f(x)) \quad (3.17)$$

where $\Sigma_\infty: \mathbb{R}^d \rightarrow \mathbb{S}_+^d$ is a map from $\mathbb{R}^d$ to the cone of symmetric positive semidefinite matrices, then one can replace the signal-to-noise part of [Assumption 3.4](#) by the following sharper requirement:

$$\liminf_{\|x\| \rightarrow \infty} \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} > 4. \quad (3.18)$$

Note that these conditions generalize [Assumption 3.3](#) and [Assumption 3.4](#): taking $\Sigma(x) = \sigma_\infty^2(f(x)) \cdot I_d$ recovers the original assumptions. These refined conditions improve the dependence on the dimension if $\Sigma_\infty(x)$ has a low effective rank i.e., its eigenvalues decay sufficiently fast so that the ratio

$\text{tr}(\Sigma_\infty(x))/\|\Sigma_\infty(x)\|_{\text{op}}$ is smaller than d . Our proofs would in that case rely tail inequalities for quadratic forms of sub-Gaussian random vectors, see e.g., (Hsu et al., 2012, Rudelson and Vershynin, 2013). 

3.2.2 Examples: regularized empirical risk minimization

In this section, we discuss some examples of settings that satisfy the above assumptions. In particular, we focus on the regularized empirical risk minimization problem:

$$f(x) = \frac{1}{N} \sum_{i=1}^N \ell(x; \xi_i) + \frac{\lambda}{2} \|x\|^2 \quad (3.19)$$

where ξ_i , are the training data of the model, $\ell(x; \xi)$ represents the loss of the model x on the data point ξ , and $\lambda > 0$ is a regularization parameter.

We consider the SFO obtained by sampling mini-batches of size B . The noise term $U(x, \omega)$ is then given by

$$U(x, \omega) = \frac{1}{B} \sum_{b=1}^B \nabla \ell(x; \xi_{\omega_b}) - \frac{1}{N} \sum_{i=1}^N \nabla \ell(x; \xi_i) \quad (3.20)$$

with $\omega = (\omega_1, \dots, \omega_B)$ representing B indices from $\{1, \dots, N\}$.

We consider two settings.

Example 3.1 (Smooth and Lipschitz losses). Let us consider the case where ℓ is C^2 -differentiable, L -Lipschitz continuous and β -smooth. These are common assumptions in non-convex optimization, – see e.g., Mertikopoulos et al. (2020) and references therein. We provide more details, showing that (3.19) with such a loss satisfies our assumptions.

Since ℓ is C^2 -differentiable and β -smooth, the objective f is $(\beta + \lambda)$ -smooth. Moreover, L -Lipschitz continuity of ℓ implies that $\ell(x; \xi) \geq \ell(0; \xi) - L\|x\|$, so that f is lower-bounded by the quadratic $\frac{\lambda}{2}\|x\|^2 - L\|x\| + \min_i \ell(0; \xi_i)$, and in particular coercive. In particular, Assumption 3.2 and (a),(b) are satisfied.

By L -Lipschitz continuity of ℓ , all the terms of the form (3.22) are uniformly bounded by $2L$, and therefore $\mathcal{O}(1)$ -sub-Gaussian (Wainwright, 2019, Ex. 2.4). By independence, $U(x, \omega)$ is $\mathcal{O}(1/B)$ -sub-Gaussian, so Assumption 3.3(b),(c) are satisfied with σ_∞^2 constant.

Finally, since $\|\nabla \ell(x; \xi)\| \leq L$ uniformly, we have $\|\nabla f(x)\| = \Omega(\lambda\|x\|)$ for $\|x\|$ large enough, and Assumption 3.4 follows.

Thus, this setting covers two wide classes of examples: linear models with non-convex losses (Erdogdu et al., 2018, Lu et al., 2021) and smooth neural networks with normalization layers² (Li et al., 2020b). 

Example 3.2 (Dissipative objectives). Our assumptions can also be linked to the notion of *dissipativity*, which is standard in Markov chain and sampling literature, see e.g., (Erdogdu et al., 2018, Li and Wang, 2022, Raginsky et al., 2017, Lu et al., 2021, Lytras and Mertikopoulos, 2024). Considering the gradient oracle obtained by sampling minibatches of size B , this setting fits into our framework provided B is chosen large enough, as detailed below.

² As noted by Li et al. (2020b), networks with a normalization layer after every linear layer, including a final one, enjoy the global Lipschitzness property.

Let us assume that ℓ is non-negative, C^2 -differentiable, β -Lipschitz smooth and that the resulting objective f is dissipative: there are $\alpha, \beta > 0$ such that

$$\langle \nabla f(x), x \rangle \geq \alpha \|x\|^2 - \Delta. \quad (3.21)$$

Then, all the terms

$$\nabla \ell(x; \xi_j) - \frac{1}{N} \sum_{i=1}^N \nabla \ell(x; \xi_i) \quad (3.22)$$

are uniformly bounded by $\mathcal{O}(\|x\|)$ by smoothness of ℓ . In particular, this implies that [Assumption 3.3\(b\)](#) is satisfied.

Moreover, we also obtain that all the terms (3.22) are $\mathcal{O}(\|x\|^2)$ -sub-Gaussian, and therefore, by independence we obtain that $U(x, \omega)$ is $\mathcal{O}(\frac{1}{B}\|x\|^2)$ -sub-Gaussian.

Since ℓ is non-negative, f is lower-bounded by $\Omega(\|x\|^2)$. It gives that $U(x, \omega)$ is actually $\mathcal{O}(\frac{1}{B}f(x))$ -sub-Gaussian and therefore [Assumption 3.3\(c\)](#) is satisfied with $\sigma_\infty^2(t) \propto \frac{t}{B}$.

We now show how [Assumption 3.4](#) can be satisfied. If the structure of the objective ensure that $U(x, \omega)$ grows at most as $\mathcal{O}(\|x\|^s)$ with $s < 2$, then [Assumption 3.4](#) is satisfied. Otherwise, if $U(x, \omega)$ actually grows as $\mathcal{O}(\|x\|^2)$ (it is the worst case since ℓ is smooth), then [Assumption 3.4](#) can still be satisfied by choosing B large enough. Indeed, by dissipativity we have that

$$\begin{aligned} \|\nabla f(x)\|^2 &= \|\nabla f(x) - \alpha x\|^2 + \alpha^2 \|x\|^2 + 2\alpha \langle \nabla f(x) - \alpha x, x \rangle \\ &\geq \alpha^2 \|x\|^2 - 2\Delta = \Omega(\|x\|^2), \end{aligned} \quad (3.23)$$

so that

$$\frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \geq \Omega\left(\frac{\|x\|^2}{f(x)/B}\right) = \Omega(B). \quad (3.24)$$

The second part of [Assumption 3.4](#) is satisfied by non-negativity and smoothness of ℓ . 

3.3 LARGE DEVIATIONS FOR SGD

As described in § 1.2, our goal is to characterize the long-run distribution of SGD and its time to reach the global minimum. The key property to control is the probability of SGD following certain trajectories over long time horizons: the invariant measure encodes how often each region is visited, while the hitting time encodes how likely SGD is to reach a given target before time T . A large deviation principle (LDP) is precisely the tool that quantifies such probabilities at the exponential scale: it characterizes, for each path γ , the probability that the iterates of SGD stay close to γ , as $\eta \rightarrow 0$. Deriving such a principle for SGD is therefore the first step of our approach. Unlike classical results ([Freidlin and Wentzell, 1998b](#), [Kifer, 1988a](#)), which concern continuous-time diffusions on compact manifolds, SGD evolves in discrete time on the unbounded space $\mathbb{R}^d$ with a general, possibly non-Gaussian noise.

3.3.1 Intuition

In our setting, a sample-path large deviation principle for SGD characterizes the probability distribution of SGD trajectories on large intervals, as the step-size η decreases. To be able to obtain a LDP, we need to consider an “accelerated”

or “subsampled” version of the SGD process by a factor $1/\eta$. This subsampled sequence essentially shares the same long-run behavior and invariant measures and, in addition, it has a very important feature: there are sufficiently many time steps between two of its iterates for an LDP to hold in the specified subinterval. Intuitively, this is because it takes $\mathcal{O}(1/\eta)$ steps of (SGD) to average out random fluctuations due to the noise.

We thus define the accelerated process x_n^η , $n = 0, 1, \dots$, by

$$x_n^\eta = x_{n\lfloor 1/\eta \rfloor} \quad (3.25)$$

where $\lfloor 1/\eta \rfloor$ denotes the integer part of $1/\eta$.

Now, given a continuous path $\gamma: [0, T] \rightarrow \mathbb{R}^d$, we want to characterize the probability that $(x_n^\eta)_{n \leq T}$ is close to γ i.e., that x_n^η is close to $\gamma(n)$ for all $n \leq T$. This is illustrated by Figure 8.

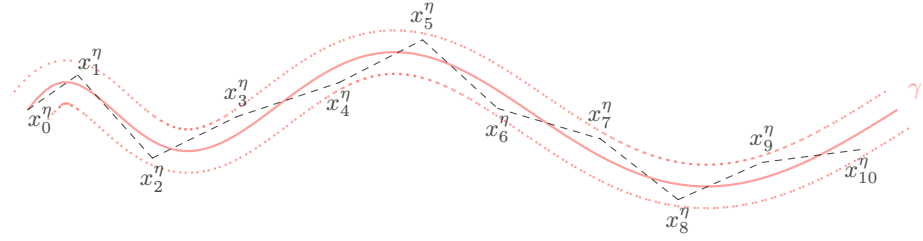


Figure 8: Illustration of the large deviation principle for SGD.

In a very informal fashion, the LDP we will rely on reads as follows:

$$\mathbb{P}\left((x_n^\eta)_{n \leq T} \approx \gamma\right) \approx \exp\left(-\frac{S_T[\gamma]}{\eta}\right) \quad (3.26)$$

where $S_T[\gamma]$ is the *action functional* applied to γ , also called the *rate function*. Therefore, this action functional fully characterizes the probability distribution of SGD trajectories on large intervals, as the step-size η decreases.

How is this action functional defined? For this, we require a few notations and definitions.

3.3.2 Hamiltonian, Lagrangian and action functional

We introduce the *Hamiltonian*, or the logarithm of the cumulant generating function of the random variable $-\nabla f(x) + U(x, \omega)$: for any $x \in \mathbb{R}^d$, $p \in \mathbb{R}^d$,

$$\mathcal{H}(x, p) := \log \mathbb{E}[\exp(\langle p, -\nabla f(x) + U(x, \omega) \rangle)] \quad (3.27)$$

$$= -\langle \nabla f(x), p \rangle + \log \mathbb{E}[\exp(\langle p, U(x, \omega) \rangle)]. \quad (3.28)$$

This function characterizes the distribution of the noise at x and therefore encompasses all the information about its statistical properties. Also notice that this function is convex in p . Its convex conjugate p is the *Lagrangian*: for any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$,

$$\mathcal{L}(x, v) := \mathcal{H}^*(x, v), \quad (3.29)$$

which plays a crucial role in large deviation theory. In particular, $\mathcal{L}(x, v)$ measures how far v is from $-\nabla f(x)$ i.e., for any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$

$$\mathcal{L}(x, v) \geq 0 \quad \text{and} \quad \mathcal{L}(x, v) = 0 \iff v = -\nabla f(x). \quad (3.30)$$

We can now introduce the “action functional” of $\mathcal{L}$ (Freidlin and Wentzell, 1998b, Chap. 3.2) as

$$\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}(\gamma(t), \dot{\gamma}(t)) dt \quad (3.31)$$

for all $\gamma: [0, T] \rightarrow \mathbb{R}^d$ absolutely continuous, where $\dot{\gamma}(t)$ denotes the derivative of γ at time t , and with the convention $\mathcal{S}_T[\gamma] = \infty$ if γ is not absolutely continuous. (3.30) implies the following properties of the action functional in relation to the flow $\Theta_t(x)$ (Definition 3.1):

Lemma 3.1 (Properties of the flow). *For any $T \geq 0$, $\gamma \in \mathcal{C}([0, T], \mathbb{R}^d)$ such that $\gamma_0 = x$,*

$$\begin{aligned} \mathcal{S}_T(\gamma) \geq 0 \quad \text{and} \quad \mathcal{S}_T(\gamma) = 0 &\iff \dot{\gamma}(t) = -\nabla f(\gamma(t)) \text{ for a.e. } t \in [0, T] \\ &\iff \gamma_t = \Theta_t(x) \text{ for a.e. } t \in [0, T]. \end{aligned} \quad (3.32)$$

Thus the action functional is non-negative and is equal to zero if and only if γ is a solution of the gradient flow. The action functional thus quantifies how much γ deviates from being a solution of the gradient flow.

To make this more concrete and continue explaining the properties of the action functional, we consider, **until the end of this section**, the special case of Gaussian noise. Note that most of these properties are actually valid, in some form, in the general case, but they are more easily explained and visualized in the Gaussian case.

Assume that the noise $U(x, \omega)$ always follows the Gaussian distribution

$$U(x, \omega) \sim \mathcal{N}(\mathbf{0}_d, \sigma^2 I_d). \quad (3.33)$$

In that case, both the Hamiltonian and the Lagrangian have simple explicit expressions:

$$\mathcal{H}(x, p) = -\langle \nabla f(x), p \rangle + \frac{\sigma^2}{2} \|p\|_2^2 \quad (3.34)$$

$$\mathcal{L}(x, v) = \frac{\sigma^2}{2} \|v + \nabla f(x)\|_2^2. \quad (3.35)$$

Thus, the action functional reads as:

$$\mathcal{S}_T[\gamma] = \frac{\sigma^2}{2} \int_0^T \|\dot{\gamma}(t) + \nabla f(\gamma(t))\|_2^2 dt. \quad (3.36)$$

This makes it clear that the action functional is a measure of how far γ is from being a solution of the gradient flow, with higher values indicating higher deviations from the gradient flow.

So, the farther γ is from being a solution of the gradient flow, the higher $\mathcal{S}_T[\gamma]$ is and, as a consequence, the less likely it is for SGD to be close to γ (3.26). Conversely, if γ is close to being a solution of the gradient flow, then $\mathcal{S}_T[\gamma]$ is low and the probability for SGD to be close to γ is not vanishingly small. However, in both cases, we still have the precise estimate of that probability, be it large or small.

Continuing with the isotropic Gaussian noise example a very informative lower-bound of $\mathcal{S}_T[\gamma]$ can be obtained: for any $\gamma \in \mathcal{C}_T$, Young’s inequality implies that

$$\frac{2(f(\gamma(T)) - f(\gamma(0)))}{\sigma^2} \quad (3.37)$$

$$\begin{aligned}
&= \frac{1}{\sigma^2} \int_0^T 2 \langle \dot{\gamma}(t), \nabla f(\gamma(t)) \rangle dt \\
&\leq \frac{1}{\sigma^2} \int_0^T \langle \dot{\gamma}(t), \nabla f(\gamma(t)) \rangle + \frac{1}{2} \|\dot{\gamma}(t)\|_2^2 + \frac{1}{2} \|\nabla f(\gamma(t))\|_2^2 dt \\
&= \frac{1}{\sigma^2} \int_0^T \frac{1}{2} \|\dot{\gamma}(t) + \nabla f(\gamma(t))\|_2^2 dt = \mathcal{S}_T[\gamma].
\end{aligned} \tag{3.38}$$

Hence, the action functional is lower-bounded by the difference of the objective values at the endpoints of γ : if $f(\gamma(T))$ is larger than $f(\gamma(0))$, then it means that γ is going “uphill” and therefore it is far from being a solution of the gradient flow and $\mathcal{S}_T[\gamma]$ is large. Note (3.38) is actually valid in the general case, not only for isotropic Gaussian noise, see § 3.6.1.

Remark 4. Though the Gaussian case is very convenient to illustrate and understand the properties of the action functional, there are a few limitations to it. The astute reader might have noticed that the Gaussian case does not formally satisfy the assumptions of our setting, since the Gaussian noise is unbounded and thus contradicts [Assumption 3.3](#). To formally apply our results to SGD with Gaussian noise, we will thus consider in the following chapters a truncated version of the Gaussian distribution. Having bounded noise is indeed necessary for a LDP to hold for our discrete-time process (this is not needed in the continuous case). Bounded noise however brings some technical challenges when working with the action functional: the Lagrangian is no longer regular, it is infinite outside of a compact set! This is one of the reasons why we need to refine the classical theories of [Freidlin and Wentzell \(1998b\)](#) and [Kifer \(1988a\)](#), which we will do in § 3.5 in particular. 

Now that we have introduced the action functional and explained its properties, we are ready to state the large deviation principle for SGD in a rigorous way.

3.3.3 Large deviation principle for “continuous-time SGD trajectories”

We start from classical LDP results for continuous-time processes ([Freidlin and Wentzell, 1998b](#), Chap. 7) and then we transfer them to the discrete-time process of interest. To do so, we introduce some continuous-time trajectories related to SGD.

From the sequences $(x_n)_{n \geq 0}$, $(\omega_n)_{n \geq 0}$, and $(x_n^\eta)_{n \geq 0}$, we define two continuous sequences (in uppercase):

- (a) The continuous “interpolated” trajectory, defined for any $n \geq 0$, $t \in [\eta n, \eta(n+1)]$ by

$$X_t = x_n + \left(\frac{t}{\eta} - n \right) (x_{n+1} - x_n) \tag{3.39a}$$

- (b) The continuous “discretized noise” trajectory, defined by $Z_0 = x_0$ and for any $t > 0$

$$\dot{Z}_t = -\nabla f(Z_t) + U(Z_t, \omega_{\lfloor t/\eta \rfloor}) \tag{3.39b}$$

First, note that both continuous-time processes are “accelerated” by a factor $1/\eta$ compared to the original sequences appearing in [\(SGD\)](#). Put simply, the first sequence X connects consecutive SGD iterates x_n and x_{n+1} with a straight line

segment on each interval $[\eta n, \eta(n+1))$. The second one, Z , is less common. The key feature of Z is that it uses the *exact same* noise samples as discrete SGD: the index $\lfloor t/\eta \rfloor$ is piecewise constant, equal to n on each interval $[\eta n, \eta(n+1))$, so the noise realization ω_n from the n -th iterate is injected into the continuous dynamics throughout that interval. Intuitively, this coupling is precisely what makes Z provably close to the discrete iterates for *every* noise realization, and not merely in distribution or in L^p -norm. This is in sharp contrast to SDE approximations of SGD, which only approximate SGD in a probabilistic sense (in distribution or in L^p -norm).

The idea is now the following: to establish a LDP for the discrete accelerated sequence $(x_n^\eta)_{n \geq 0}$, we build upon [Freidlin and Wentzell \(1998b, Chap. 7\)](#) to obtain a LDP on $(Z_t)_{t \geq 0}$ then we transfer it to $(X_t)_{t \geq 0}$, and finally we deduce the result for $(x_n^\eta)_{n \geq 0}$.

To do so, we begin by equipping the space of continuous functions $\mathcal{C}([0, T], \mathbb{R}^d)$ with the distance induced by the uniform norm on $[0, T]$ for any $T > 0$: for any $\gamma, \varphi \in \mathcal{C}([0, T], \mathbb{R}^d)$,

$$\text{dist}_{[0, T]}(\gamma, \varphi) := \sup_{t \in [0, T]} \|\gamma_t - \varphi_t\|. \quad (3.40)$$

One can show that the distance between the (continuous) “interpolated” trajectory and the “discretized noise” trajectory is bounded by $\mathcal{O}(\eta)$ almost surely on finite intervals. We will consider the set of trajectories with bounded action functional:

$$\Gamma_{[0, T]}^{\mathcal{K}}(s) := \{\gamma \in \mathcal{C}([0, T], \mathbb{R}^d) : \gamma_0 \in \mathcal{K}, \mathcal{S}_{[0, T]}(\gamma) \leq s\} \quad (3.41)$$

From [Freidlin and Wentzell \(1998b, Chap. 7\)](#) we obtain a large deviations principle on $(Z_t)_{t \geq 0}$ and then transfer it to $(X_t)_{t \geq 0}$

Proposition 3.2. *Fix some $T > 0$, some $\mathcal{K} \subset \mathbb{R}^d$ compact, and some $s > 0$.*

- *The set $\Gamma_{[0, T]}^{\mathcal{K}}(s)$ is compact and $\mathcal{S}_{[0, T]}$ is lower semi-continuous (l.s.c.) on $\mathcal{C}([0, T], \mathbb{R}^d)$.*
- *For any $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathbb{R}^d$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, we have that*

$$\mathbb{P}_{x_0}(\text{dist}_{[0, T]}(X, \gamma) < \delta) \geq \exp\left(-\frac{\mathcal{S}_{[0, T]}(\gamma) + \varepsilon}{\eta}\right) \quad (3.42a)$$

$$\mathbb{P}_{x_0}(\text{dist}_{[0, T]}(X, \Gamma_{[0, T]}^{\{x_0\}}(s)) > \delta) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right) \quad (3.42b)$$

for all $\gamma \in \Gamma_{[0, T]}^{\{x_0\}}(s)$.

Taken together, the two statements of [Proposition 3.2](#) are a formal version of the informal LDP stated in (3.26) for the continuous-time process $(X_t)_{t \geq 0}$.

3.3.4 Large deviation principle for accelerated SGD

We now leverage the discussion in the previous section to establish a LDP for the discrete accelerated trajectories $(x_n^\eta)_{n \geq 0}$ defined in (3.25).

To do so, we need a discrete analogue of the continuous action functional $\mathcal{S}_T$ from (3.31). The central object is the *single-step rate function*: for any $x, x' \in \mathbb{R}^d$,

$$\rho(x, x') := \inf\{\mathcal{S}_{0,1}(\gamma) : \gamma \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}^d), \gamma_0 = x, \gamma_1 = x'\}. \quad (3.43)$$

Thus $\rho(x, x')$ represents the minimum cost, measured by the continuous action functional over a unit time interval, to travel from x to x' . Like for the action functional, we have the following relationship between the single-step rate function and the flow $\Theta_t(x)$:

Lemma 3.3 (Properties of ρ and F). *For any $x, x' \in \mathbb{R}^d$,*

$$\rho(x, x') = 0 \iff x' = \Theta_1(x). \quad (3.44)$$

In a very informal fashion, the single-step LDP reads as follows: for any $x, x' \in \mathbb{R}^d$,

$$\mathbb{P}_x(x_1^\eta \approx x') \sim \exp\left(-\frac{\rho(x, x')}{\eta}\right). \quad (3.45)$$

We then define, for $N \geq 0$ and $\xi = (\xi_0, \dots, \xi_{N-1}) \in (\mathbb{R}^d)^N$, the *discrete action functional*

$$\mathcal{S}_N(\xi) := \sum_{n=0}^{N-2} \rho(\xi_n, \xi_{n+1}). \quad (3.46)$$

Summing the single-step costs along a path, $\mathcal{S}_N(\xi)$ measures how costly it is for the accelerated process to follow the discrete trajectory ξ . In a very informal fashion, the multi-step LDP reads as follows:

$$\mathbb{P}_x\left((x_n^\eta)_{0 \leq n \leq N-1} \approx \xi\right) \approx \exp\left(-\frac{\mathcal{S}_N(\xi)}{\eta}\right). \quad (3.47)$$

where $\mathcal{S}_N(\xi)$ is the discrete action functional defined in (3.46). The following proposition [Proposition 3.4](#) makes this precise. We begin by introducing the necessary notations and definitions. For any $N \geq 0$ and any sequences $\xi_n, \zeta_n \in \mathbb{R}^d$, $n = 0, \dots, N-1$, we define the distance:

$$\text{dist}_N(\xi, \zeta) = \max_{0 \leq n \leq N-1} \|\xi_n - \zeta_n\|, \quad (3.48)$$

and the set of discrete trajectories with low action functional

$$\Gamma_N^{\mathcal{K}}(s) := \{\xi \in (\mathbb{R}^d)^N : \xi_0 \in \mathcal{K}, \mathcal{S}_N(\xi) \leq s\} \quad (3.49)$$

Proposition 3.4. *Fix some $N \geq 0$, some $\mathcal{K} \subset \mathbb{R}^d$ compact, and some $s > 0$. Then:*

- *The set $\Gamma_N^{\mathcal{K}}(s)$ is compact and $\mathcal{S}_N$ is l.s.c. on $(\mathbb{R}^d)^N$.*
- *For all $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathbb{R}^d$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for all $x_0 \in \mathcal{K}$, $n \leq N$, $\xi \in \Gamma_n^{\{x_0\}}(s)$, we have that*

$$\mathbb{P}_{x_0}(\text{dist}_n(x^\eta, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{S}_n(\xi) + \varepsilon}{\eta}\right) \quad (3.50a)$$

and

$$\mathbb{P}_{x_0}(\text{dist}_n(x^\eta, \Gamma_n^{\{x_0\}}(s)) > \delta) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \quad (3.50b)$$

In this result, the functional $\eta^{-1}\mathcal{S}_N$ is thus the action functional in $(\mathbb{R}^d)^N$ of the process $(x_n^\eta)_{0 \leq n \leq N-1}$ uniformly with respect to the starting point x_0 in any compact set $\mathcal{K} \subset \mathbb{R}^d$, as $\eta \rightarrow 0$. The proof of [Proposition 3.4](#) relies on the large deviation principle for the continuous-time process $(X_t)_{t \geq 0}$ established in [Proposition 3.2](#) and on the fact that the discrete accelerated iterates $(x_n^\eta)_n$ are close

to the continuous interpolated trajectory $(X_t)_{t \geq 0}$ at times $n = 0, \dots, N - 1$ with an error of order $\mathcal{O}(\eta)$ almost surely. The proof can be found in [Appendix A.2](#).

The single-step rate function ρ defined in (3.43) is the key object to understand the large deviation principle for the discrete iterates $(x_n^\eta)_n$.

3.4 THE INDUCED CHAIN

We now define the induced chain $(z_n)_{n \geq 0}$ mentioned in § 3.1.3: we will show that this chain captures the long-run behaviour of SGD and that it can be treated as a Markov chain on the finite set of critical components $\{\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}\}$, allowing us to derive quantitative estimates.

Consider $\mathcal{C}_i^\delta$, $i = 1, \dots, N_{\text{crit}}$ disjoint δ -neighborhoods of $\mathcal{C}_i$, i.e., for any $i \in \{1, \dots, N_{\text{crit}}\}$,

$$\mathcal{C}_i^\delta := \mathcal{U}_\delta(\mathcal{C}_i) = \{x : d(x, \mathcal{C}_i) < \delta\} \quad (3.51)$$

for δ small enough. Denote by $\mathcal{C}^\delta := \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$ their union. Put simply, the induced chain is obtained by discarding all the iterates of $(x_n^\eta)_{n \geq 0}$ that are not in $\mathcal{C}^\delta$ and keeping only the iterates that are in $\mathcal{C}^\delta$; [Figure 9](#) illustrates this construction.

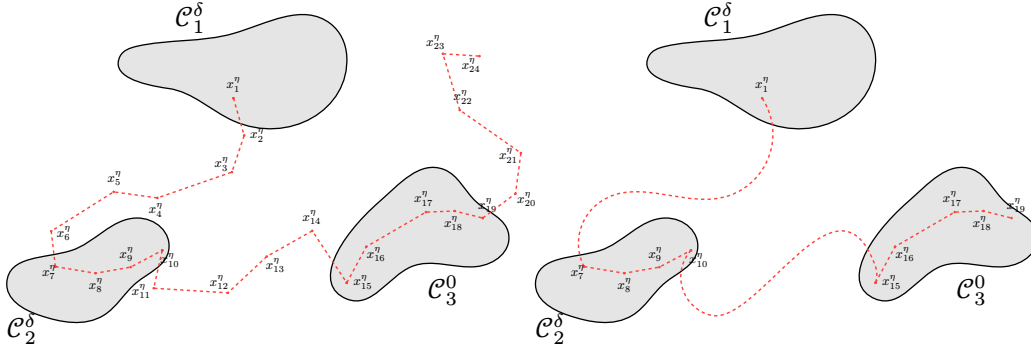


Figure 9: Illustration of the construction of the induced chain: left, the full trajectory of SGD; right, the trajectory of the induced chain restricted to iterates inside the components.

Let us define an important object: the law of the (accelerated) iterates at the first time they reach some set $\mathcal{C}^\delta$ (typically a neighborhood of the critical set), following e.g., [Kifer \(1988a, Prop. 5.3\)](#) and [Douc et al. \(2018, Chap. 3.4\)](#).

Definition 3.5 ([Kifer \(1988a, Prop. 5.3\)](#)). For $\mathcal{C}^\delta$ defined above, we define recursively, for $n \geq 0$,

$$\tau_{n+1}^\eta := \inf\{k > \tau_n^\eta : x_k^\eta \in \mathcal{C}^\delta\} \quad (3.52)$$

with $\tau_0^\eta := 0$. We then define the induced chain $(z_n)_{n \geq 0}$ as $z_n := x_{\tau_n^\eta}^\eta$ for $n \geq 0$. For $x \in \mathbb{R}^d$, let $\mathbf{Q}(x, \cdot)$ be the law of x_1^η started at x and $\mathbf{Q}^N(x, \cdot)$ be the N -step transition probability distribution of x_N^η started at x .

In words, $\mathbf{Q}(x, \cdot)$ is the distributions of the x^η started at x at the first time they reach $\mathcal{C}^\delta$. Typically, $\mathbf{Q}(x, \mathcal{U})$ is the probability that $\mathcal{U} \subset \mathcal{C}^\delta$ is reached first among $\mathcal{C}^\delta$. Then, $\mathbf{Q}^N(x, \cdot)$ is the distribution of the x^η started at x at the N -th time they reach $\mathcal{C}^\delta$.

Remark 5. The definition above will need to be refined in [Chapter 5](#). The issue is the following: if x_n^η enters $\mathcal{C}_i^\delta$ at time n and remains there at time $n + 1$, then

both x_n^η and x_{n+1}^η are recorded as iterates of the induced chain. This creates ambiguity when relating the hitting times of the induced chain back to those of the original chain (x_n^η) : it is unclear which of the two iterates should be considered as the “arrival” in $\mathcal{C}_i^\delta$. To resolve this, we will redefine the induced chain so that only the *first* iterate upon each entry into $\mathcal{C}_i^\delta$ is retained, yielding a one-to-one correspondence between entries of the two chains. $\heartsuit$

The induced chain $(x_n^\eta)_{n \geq 0}$ is the crucial tool in our analysis of the long-run behavior of SGD. We now carry out the three-step program outlined in § 3.1 which we recall here, along with the sections in which they are detailed:

STEP 1: REDUCTION TO A FINITE-STATE SPACE (§ 3.5). We will show that it is possible to abstract away what happens inside the critical components themselves. Instead of treating $(x_n^\eta)_{n \geq 0}$ as a Markov chain on $\mathbb{R}^d$, we will show that it can be treated as a Markov chain on the finite set of critical components. More precisely, we will show that, the behavior of SGD started at $x \in \mathcal{C}_i$ is roughly similar to the behavior of SGD started at another point x' in the same critical component $\mathcal{C}_i$. This will be done by carefully studying the regularity of the Lagrangian and the action functional near the critical components and leveraging the structure of SGD as a gradient system

STEP 2: ACCURATE MODELING BY THE INDUCED CHAIN (§ 3.6). One of the main challenges is to show that the induced chain actually represents the long-run behavior of the full dynamics. For this, we need to show, on long time horizons, that the iterates of (SGD) spend most of their time near the critical components. This requires showing that SGD does not spend much time at infinity. This will also yields bounds on the transition times in between neighborhoods of critical components.

STEP 3: TRANSITION PROBABILITIES (§ 3.7). The last key ingredient is obtaining precise estimates on the transition probabilities of the induced chain: if SGD is started in some $\mathcal{C}_i$, how likely is it to reach some $\mathcal{C}_j$? This will be done by leveraging the large deviations estimates (obtained in § 3.3).

3.5 REDUCTION TO A FINITE STATE-SPACE

The accelerated chain $(x_n^\eta)_{n \geq 0}$ is a Markov chain taking values in $\mathcal{C}^\delta = \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$, which is a subset of $\mathbb{R}^d$. The goal of this section is to show that one can abstract away what happens inside the critical components and treat $(x_n^\eta)_{n \geq 0}$ as a Markov chain on the finite set of critical components $\{\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}\}$. This is a key step to leverage explicit formulas for the asymptotic distribution of finite-state Markov chains.

The main results of this section are [Lemma 3.9](#) and [Lemma 3.10](#). Together, they justify treating the induced chain as a Markov chain on the finite set $\{\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}\}$ by showing that the starting point within a critical component is irrelevant to leading order. More precisely, [Lemma 3.9](#) shows that any two points x, x' in the same critical component $\mathcal{C}_i$ can be connected by a path of arbitrarily small cost that remains close to $\mathcal{C}_i$: in particular, SGD can travel between x and x' without paying any exponential cost and without leaving $\mathcal{C}_i$. [Lemma 3.10](#) then extends this to the neighborhoods $\mathcal{C}_i^\delta$ in which the induced chain actually takes its values: it shows that these neighborhoods can be chosen so that the cost of moving from any point in $\mathcal{C}_i^\delta$ to some point in $\mathcal{C}_i$, and back, is arbitrarily small. Hence, for the purpose of computing transition costs between

components, the precise location of the iterates within $\mathcal{C}_i^\delta$ is immaterial. The argument proceeds in three steps.

First, in § 3.5.1, we introduce the quasi-potential $\mathcal{B}$ following Kifer (1988a) and define the equivalence relation $\sim$ on $\text{crit } f$. We then show that the equivalence classes of $\sim$ coincide with the connected components of $\text{crit } f$ (Corollary 3.8): two points can be connected by a zero-cost path in both directions if and only if they belong to the same connected component of $\text{crit } f$.

Second, in § 3.5.2, we establish the regularity properties of $\mathcal{L}$ and ρ near critical points of f that are needed to prove the above. In the classical setting of Freidlin and Wentzell (1998b) and Kifer (1988a), where the noise is Gaussian and the state space is a compact manifold, such regularity comes for free; here, since the noise is bounded, it must be carefully established by exploiting the gradient structure of the problem (Lemmas 3.6 and 3.7).

Third, in § 3.5.3 and § 3.5.4, we combine these ingredients to prove the main results. Lemma 3.9 shows that any two points in the same critical component can be connected by a low-cost path that stays close to that component; Lemma 3.10 constructs neighborhoods of critical components with arbitrarily small entry and exit costs, which is needed to reduce the induced chain — whose iterates lie in neighborhoods of the critical components, not in the components themselves — to a finite-state Markov chain.

3.5.1 Equivalence classes of points in $\text{crit } f$

Let us first formalize the minimum-energy displacement between two points, which we refer to as the *quasi-potential* between those points.

Definition 3.6 (Kifer (1988a, §1.5)). Define, for $x, x' \in \mathbb{R}^d$,

$$\begin{aligned} \mathcal{B}(x, x') &= \inf \{ \mathcal{S}_{[0,T]}(\gamma) : \gamma \in \mathcal{C}([0, T], \mathbb{R}^d), \gamma_0 = x, \gamma_T = x', T \in \mathbb{N}, T \geq 1 \} \\ &= \inf \{ \mathcal{S}_N(\xi) : \xi \in (\mathbb{R}^d)^N, \xi_0 = x, \xi_{N-1} = x', N \geq 1 \}. \end{aligned} \quad (3.53)$$

The fact that these two expressions coincide directly come from the definition of ρ . $\mathcal{B}(x, x')$ represents the minimum cost to go from x to x' in any number of steps and thus how likely it is for SGD to go from x to x' in any number of steps: the LDP from § 3.3 implies that, for N large enough,

$$\mathbb{P}_x(x_{N-1}^\eta \approx x') \approx \exp\left(-\frac{\mathcal{B}(x, x')}{\eta}\right). \quad (3.54)$$

Therefore, the higher $\mathcal{B}(x, x')$, the less likely it is for SGD to go from x to x' in any number of steps. Conversely, if $\mathcal{B}(x, x')$ is zero, then it is not exponentially unlikely for SGD to go from x to x' in any number of steps.

Note that $\mathcal{B}$ inherits the properties of $\mathcal{S}$ detailed in § 3.3.1. In particular, if there exists a solution of the gradient flow going from x to x' , i.e., a curve $\gamma: [0, T] \rightarrow \mathbb{R}^d$ such that

$$\gamma_0 = x, \quad \gamma_T = x' \quad \text{and} \quad \dot{\gamma}_t = -\nabla f(\gamma_t) \text{ for all } t \in [0, T], \quad (3.55)$$

then $\mathcal{B}(x, x') = 0$. Hence, SGD is very likely to transition from x to x' .

Conversely, in the case of the Gaussian noise § 3.3.2, (3.38) implies that $\mathcal{B}(x, x')$ is lower-bounded as:

$$\mathcal{B}(x, x') \geq \frac{2(f(x') - f(x))}{\sigma^2}. \quad (3.56)$$

Hence, if $f(x')$ is much larger than $f(x)$, then $\mathcal{B}(x, x')$ is large and it is very unlikely for SGD to go from x to x' : SGD has to “go uphill” to go from x to x' , which is unlikely. Nevertheless, in both cases, (3.54) still provides us with an estimate of that probability.

Following Kifer (1988a) (see also Freidlin and Wentzell (1998b) in a continuous-time setting), we now define a crucial equivalence relation between points in $\text{crit } f$ based on the quasi-potential. Intuitively, two points $x, x' \in \text{crit } f$ are equivalent if they can be connected by a path of zero cost, which means that it is possible to move from x to x' and from x' to x without incurring any cost.

Proposition 3.5 (Kifer (1988a, §1.5)). *The relation $\sim$ defined for any $x, x' \in \text{crit } f$ as*

$$x \sim x' \iff \mathcal{B}(x, x') = \mathcal{B}(x', x) = 0 \quad (3.57)$$

is an equivalence relation on $\text{crit } f$.

The equivalence classes of $\sim$ play a central role in the theories of Kifer (1988a) and Freidlin and Wentzell (1998b).

We will show that, in our setting, the equivalence classes of $\sim$ coincide with the connected components of $\text{crit } f$. However, to do so, we first need to establish some regularity properties of the Lagrangian $\mathcal{L}$ and the rate function ρ near critical points of f .

3.5.2 Regularity of the Lagrangian and the rate function near critical points

In the classical setting of Freidlin and Wentzell (1998b) and Kifer (1988a), where the noise is Gaussian and the state space is a compact manifold, the rate function ρ is uniformly continuous. This is far from being the case here: since the noise is bounded, $\mathcal{L}$ has a bounded domain, and $\rho(x, x')$ may be infinite for x' far from x . Despite this, we show that some continuity can be salvaged near critical points of f — and this is precisely what is needed for the reduction to a finite-state space.

Lemma 3.6. *For any $x \in \text{crit}(f)$, $\rho(x, x) = 0$ and, as $x' \rightarrow x$, $\rho(x, x') \rightarrow 0$ and $\rho(x', x) \rightarrow 0$.*

This continuity property crucially uses that, in gradient systems, all solutions of the problem are zeros of the vector field, which is not the case in general dynamical systems; in particular, this approach does not extend to non-gradient systems. It is a consequence of the following regularity property of $\mathcal{L}$ near critical points of f .

Recall that $\mathcal{L}(x, v)$ measures the cost for SGD to move in direction v from x : it vanishes at $v = -\nabla f(x)$ and is positive otherwise. The regularity of $\mathcal{L}$ near $v = \mathbf{0}_d$ requires two conditions: first, $\nabla f(x) = \mathbf{0}_d$, so that the zero-cost direction is $\mathbf{0}_d$ itself; second, the noise covariance $\text{cov}(U(x; \omega))$ must be positive definite, so that SGD can move in every direction from x . At critical points of f , where $\nabla f(x) = \mathbf{0}_d$, both conditions are satisfied and the implicit function theorem yields the desired regularity, meaning in particular that SGD can move in any direction at a cost that is at most quadratic in the norm of the direction.

Lemma 3.7. *For any $x \in \mathbb{R}^d$, there exists $\delta > 0$ such that $(x', v) \mapsto \mathcal{L}(x', v + \nabla f(x'))$ is finite and jointly Lipschitz continuous on $\mathbb{B}(x, \delta) \times \mathbb{B}(0, \delta)$. Moreover, there exists a constant $c > 0$ such that, for all $x' \in \mathbb{B}(x, \delta) \cap \text{crit } f$ and all $v \in \mathbb{B}(0, \delta)$,*

$$\mathcal{L}(x', v) \leq c \|v\|^2. \quad (3.58)$$

This quadratic upper bound on the cost of moving away from a critical point is the key to [Lemma 3.6](#): since $\rho(x, x')$ is obtained by optimizing over one-step paths from x to x' , and since the cost of a single step of size $\|x' - x\|$ is at most $c\|x' - x\|^2$ near a critical point, $\rho(x, x')$ vanishes as $x' \rightarrow x$.

3.5.3 Equivalence classes coincide with connected components of $\text{crit } f$

There are three main ingredients to show that the equivalence classes of $\sim$ coincide with the connected components of $\text{crit } f$:

- The regularity properties above ([Lemma 3.6](#)) show that the equivalence classes of $\sim$ are closed.
- By construction, the equivalence classes of $\sim$ are connected.
- The final step consists in leveraging the connectedness requirement on $\text{crit}(f)$ ([Assumption 3.2](#)): we obtain that any connected component of $\text{crit } f$ is included in an equivalence class of $\sim$.

Putting everything together, by definition of connected components, this means the equivalence classes of $\sim$ coincide exactly with the connected components of $\text{crit } f$.

Corollary 3.8. *The equivalence classes of $\sim$ coincide with the connected components of $\text{crit } f$.*

In particular, this means that f is constant on each equivalence class of $\sim$: if $x \sim x'$, then $f(x) = f(x')$. This result has two important consequences:

- If $x \sim x'$, i.e.,

$$\mathcal{B}(x, x') = \mathcal{B}(x', x) = 0, \quad (3.59)$$

then x and x' are in the same connected component of $\text{crit } f$. This means that there no “cycles” of zero cost transitions between different connected components of $\text{crit } f$: if x and x' are in different connected components of $\text{crit } f$, then either $\mathcal{B}(x, x') > 0$ or $\mathcal{B}(x', x) > 0$.

- If x and x' are in the same connected component of $\text{crit } f$, then $\mathcal{B}(x, x') = \mathcal{B}(x', x) = 0$. This means that it is not exponentially unlikely for SGD to go from x to x' and from x' to x : for N, N' large enough,

$$\mathbb{P}_x(x_{N-1}^\eta \approx x') \geq \exp(-\varepsilon/\eta) \quad \text{and} \quad \mathbb{P}'_x(x_{N'-1}^\eta \approx x) \geq \exp(-\varepsilon/\eta). \quad (3.60)$$

The next section will provide a stronger version of this result, in particular guaranteeing that the trajectory connecting x and x' remains close to the connected component of $\text{crit } f$ containing x and x' .

3.5.4 Putting everything together

We can now present the main results of this section. To show that the behavior of SGD started at x is roughly similar to the behavior of SGD started at x' if x and x' are in the same critical component $\mathcal{C}_i$, we show that any such x and x' can be connected by a path of low cost that remains close to the equivalence class of x and x' with uniformly bounded length (see [Figure 10](#)). This is formalized in the following lemma.

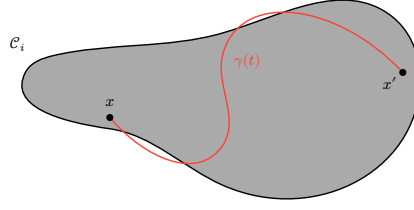


Figure 10: Illustration of the construction of a low-cost path between two points x, x' in the same equivalence class $\mathcal{K}$. The red curve represents the low-cost path connecting x and x' that remains close to $\mathcal{K}$.

Lemma 3.9. *Let $\mathcal{K}$ be an equivalence class of $\sim$. Then, for any $\varepsilon > 0$, there is some $N \geq 1$ such that, for any $x, z \in \mathcal{K}$, there is $\xi \in (\mathbb{R}^d)^N$ such that $\xi_0 = x$, $\xi_{N-1} = z$, $\mathcal{S}_N(\xi) < \varepsilon$ and $\max_{0 \leq n < N} d(\xi_n, \mathcal{K}) < \varepsilon$.*

Combined with the LDP estimates obtained previously, this result should be understood as follows: for all $\mathcal{C}_i$ connected component of $\text{crit } f$ and all $\varepsilon > 0$, there is $N \geq 1$ such that the probability that SGD started at $x \in \mathcal{C}_i$ reaches some $x' \in \mathcal{C}_i$ in N steps without ever leaving a ε -neighborhood of $\mathcal{C}_i$ is not exponentially small in η :

$$\mathbb{P}_x \left(x_{N-1}^\eta \approx x', d(x_n^\eta, \mathcal{C}_i) < \varepsilon \text{ for all } n = 0, \dots, N-1 \right) \geq \exp(-\varepsilon/\eta) \quad (3.61)$$

This lemma is significantly expanded from Kifer (1988a, §1.5, Lem. 5.2) — which is designed for SDE on compact manifolds — to handle both the unboundedness of the space and the fact that $\mathcal{B}$ is neither l.s.c. nor upper semi-continuous (u.s.c.). We remedy both these issues by leveraging the gradient structure of the problem. Indeed, the key point in the proof of Lemma 3.9 is guaranteeing that the trajectory connecting x and x' remains close to the equivalence class $\mathcal{K}$. The proof goes as follows: since $\mathcal{K}$ is a connected component of $\text{crit}(f)$, all the points of $\mathcal{K}$ share the same objective value $f(\mathcal{K}) = f(x) = f(x')$. If γ is a continuous path over $[0, T]$ connecting x and x' with negligible action cost $\mathcal{S}_T(\gamma) \leq \varepsilon$, then (3.38) implies that, for all $t \in [0, T]$,

$$f(\gamma_t) - f(x) \leq \frac{\sigma^2 \varepsilon}{2}, \quad \text{and} \quad f(x') - f(\gamma_t) \leq \frac{\sigma^2 \varepsilon}{2}, \quad (3.62)$$

so that $f(\gamma_t)$ remains close to $f(\mathcal{K})$ for all $t \in [0, T]$. Since $\mathcal{K}$ is a connected component of $\text{crit } f$, f cannot remain constant along γ if γ leaves a neighborhood of $\mathcal{K}$. This implies that γ must remain close to $\mathcal{K}$ to have a low cost.

In the definition of the induced chain (Definition 3.5), we consider iterates that lie in neighborhoods of the critical components, not necessarily in the critical components themselves. To be able to reduce the iterates of the induced chain to the component they are close to, we also need to ensure that the behavior of SGD departing from (resp. arriving at) a neighborhood of a critical component is similar to the behavior of SGD departing from (resp. arriving at) the critical component in question.

This is achieved by the following lemma which is a consequence of the regularity properties of the rate function ρ near critical points of f established above (Lemma 3.6). It shows that there are no “barriers” of positive cost between points in a neighborhood of a critical component and points in the critical component itself. More precisely, it constructs neighborhoods of critical components such that the cost to go from any point in the neighborhood to some point in the critical component in one step is small, and conversely.

Lemma 3.10. *For any set $\mathcal{C} \subset \text{crit } f$ and $x \in \mathbb{R}^d$, define*

$$\rho(x, \mathcal{C}) := \inf_{x' \in \mathcal{C}} \rho(x, x') \quad \text{and} \quad \rho(\mathcal{C}, x) := \inf_{x' \in \mathcal{C}} \rho(x', x). \quad (3.63)$$

Then, for any $\mathcal{C} \subset \text{crit } f$ and $r > 0$, the set

$$\mathcal{W}_r(\mathcal{C}) := \{x \in \mathbb{R}^d : \rho(x, \mathcal{C}) < r, \rho(\mathcal{C}, x) < r\} \quad (3.64)$$

is a neighborhood of $\mathcal{C}$.

This lemma will allow us to consider hitting times to neighborhoods of critical components while defining transition costs that are independent of the particular choice of neighborhoods of critical components.

3.6 TRANSITION TIMES

The main focus of this section is estimating the transition times between neighborhoods of critical components: if SGD is started in close to $\text{crit}(f)$, how long does it take to come back to some neighborhood of $\text{crit}(f)$?

The main result is the following lemma that bounds $\tau_{\mathcal{U}}^\eta$ the return time to $\mathcal{U}$ a neighborhood of $\text{crit } f$. More precisely, we obtain exponential tail bounds on the return time to $\mathcal{U}$, that we denote $\tau_{\mathcal{U}}^\eta$, and bounds on its exponential moments (see Figure 11 for an illustration of the setting).

Proposition 3.11. *Let $\mathcal{U}$ be a neighborhood of $\text{crit } f$ and let $\mathcal{X}$ be a compact set such that $\mathcal{U} \subset \mathcal{X}$. Then, there is some $a, b > 0$ such that, for η small enough, for any $n \geq 0$ and any $x \in \mathcal{X}$,*

$$\mathbb{P}_x(\tau_{\mathcal{U}}^\eta > n) \leq \exp\left(-\frac{an}{\eta} + b\right), \quad (3.65)$$

or, equivalently, there is some $a, b > 0$ such that, for $\eta, \alpha > 0$ small enough, for any $x \in \mathcal{X}$,

$$\mathbb{E}_x \left[e^{\frac{\alpha \tau_{\mathcal{U}}^\eta}{\eta}} \right] \leq e^{\frac{a\alpha}{\eta} + b}. \quad (3.66)$$

This lemma guarantees us a fine control on the return time to $\mathcal{U}$: we not only have a bound on its expectation but also on its tail and exponential moments. Proposition 3.11 is obtained by combining “classical” optimization tools — Lyapunov analysis, subGaussian concentration inequalities — and large deviations estimates.

We first state the formal definition of the hitting and exit times that we will be estimating in this section.

Definition 3.7 (Stopping times for the accelerated process). For any set $A \subset \mathbb{R}^d$, we define the hitting and exit times of A :

$$\tau_A^\eta := \inf\{n \geq 1 : x_n^\eta \in A\} \quad (3.67a)$$

$$\sigma_A^\eta := \inf\{n \geq 0 : x_n^\eta \notin A\}. \quad (3.67b)$$

There are two main ingredients to the proof of Proposition 3.11: a Lyapunov analysis at infinity and a large deviation-based analysis near the critical region. Note that, in contrast to the previous section, the approach and all but one of the results of this section are entirely original.

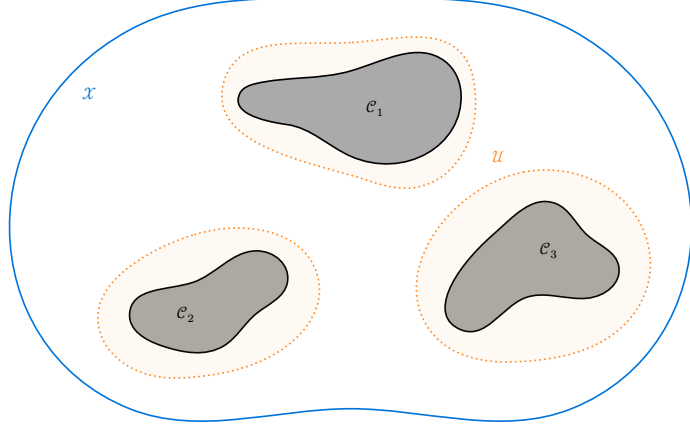


Figure 11: Illustration of the setting of [Proposition 3.11](#) with $\text{crit}(f) = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$.

3.6.1 Lyapunov potential

We now introduce a Lyapunov potential which is defined using σ_∞^2 the function that controls the subGaussianity of the noise in our assumptions [Assumption 3.3\(c\)](#).

Definition 3.8 (Lyapunov Potential). Define, for $x \in \mathbb{R}^d$

$$V_\infty(x) = 2\alpha_\infty(f(x)) \quad (3.68)$$

where $\alpha_\infty : \mathbb{R} \rightarrow \mathbb{R}$ is a primitive of $1/\sigma_\infty^2$:

$$\alpha'_\infty(t) = \frac{1}{\sigma_\infty^2(t)} \quad \text{for all } t \in \mathbb{R}. \quad (3.69)$$

If σ_∞^2 is constant, then V_∞ is just a rescaling of f . But if it is not the case, V_∞ can grow slower than f at infinity. V_∞ plays a crucial role in both the discrete-time Lyapunov analysis and the large deviations analysis. The Lyapunov analysis at infinity relies on the following descent lemma:

Lemma 3.12 (Lyapunov condition). *There exists some $\mathcal{K} \subset \mathbb{R}^d$ compact, such that, for η small enough, for any $n \geq 0$, almost surely*

$$x_n \notin \mathcal{K} \implies V_\infty(x_{n+1}) - V_\infty(x_n) \leq \eta \left(\frac{\|U(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} \right). \quad (3.70)$$

This lemma is obtained through a standard descent lemma analysis combined with our assumptions [Assumption 3.3](#). This inequality highlights the role of the SNR $\|\nabla f(x)\|^2 / \sigma_\infty^2(f(x))$: if this ratio is large enough, then V_∞ is expected to decrease at each step. In particular, [Assumption 3.4](#) ensures that this ratio becomes greater than a well-chosen constant at infinity. Iterating [Lemma 3.12](#) and using a subGaussian concentration inequality to control the sum of the squared noise terms, we then obtain the following subexponential tail bound on the return time to $\mathcal{X}$:

Lemma 3.13. *There exists a compact set $\mathcal{X} \subset \mathbb{R}^d$, such that for any larger compact set $\mathcal{X}' \supset \mathcal{X}$, there exists $a, b > 0$, such that, for η small enough, for any $x \in \mathcal{X}'$ and any $n \geq 0$,*

$$\mathbb{P}_x(\tau_{\mathcal{X}}^\eta > n) \leq \exp\left(-\frac{an}{\eta} + b\right), \quad (3.71)$$

or, equivalently, there exist $a, b > 0$ such that, for any $\eta, \alpha > 0$ small enough, $x \in \mathcal{X}'$,

$$\mathbb{E}_x \left[e^{\frac{\alpha \tau_{\mathcal{X}}^\eta}{\eta}} \right] \leq e^{\frac{a\alpha}{\eta} + b}. \quad (3.72)$$

In this lemma, $\mathcal{X}$ is a – potentially very large – set of interest: we wish to bound the return time to $\mathcal{X}$ for points outside of $\mathcal{X}$. Our bounds are uniform over all points in $\mathcal{X}'$, which can be any compact set containing $\mathcal{X}$.

V_∞ is not only useful for Lemma 3.12 and Lemma 3.13: it also plays a crucial role in the large deviations analysis near the critical region due to its close link to the action functional $\mathcal{S}_T$. We shall now outline this relationship, which the generalization of (3.38).

Using the definition of $\mathcal{H}$, Assumption 3.3(c) can be rewritten as: for any $x \in \mathbb{R}^d$ and $p \in \mathbb{R}^d$,

$$\mathcal{H}(x, p) = -\langle \nabla f(x), p \rangle + \log \mathbb{E}[\exp(\langle p, U(x, \omega) \rangle)] \quad (3.73)$$

$$\leq -\langle \nabla f(x), p \rangle + \frac{\sigma_\infty^2(f(x)) \|p\|^2}{2}. \quad (3.74)$$

Taking the convex conjugate of $\mathcal{H}$ with respect to p , we thus obtain that the Lagrangian $\mathcal{L}$ satisfies, for any $x, v \in \mathbb{R}^d$,

$$\mathcal{L}(x, v) \geq \frac{\|v + \nabla f(x)\|^2}{2\sigma_\infty^2(f(x))}. \quad (3.75)$$

We can now proceed as in (3.38): for any $\gamma \in \mathcal{C}_T$, the definition of V_∞ Definition 3.8 and Young's inequality implies that

$$\begin{aligned} V_\infty(\gamma(T)) - V_\infty(\gamma(0)) &= \int_0^T \frac{2\langle \dot{\gamma}(t), \nabla f(\gamma(t)) \rangle}{\sigma_\infty^2(f(\gamma(t)))} dt \\ &\leq \int_0^T \frac{\langle \dot{\gamma}(t), \nabla f(\gamma(t)) \rangle + \frac{1}{2} \|\dot{\gamma}(t)\|_2^2 + \frac{1}{2} \|\nabla f(\gamma(t))\|_2^2}{\sigma_\infty^2(f(\gamma(t)))} dt \\ &= \int_0^T \frac{\|\dot{\gamma}(t) + \nabla f(\gamma(t))\|_2^2}{2\sigma_\infty^2(f(\gamma(t)))} dt \\ &\leq \int_0^T \mathcal{L}(\gamma(t), \dot{\gamma}(t)) dt = \mathcal{S}_T(\gamma), \end{aligned} \quad (3.76)$$

where the last inequality follows from (3.75). A crucial consequence is that any path that makes V_∞ increase by a large amount must have a large action cost and therefore be unlikely to be followed by SGD.

3.6.2 Analysis near the critical region

In the previous subsection, we showed that SGD does not spend much time outside of $\mathcal{X}$ from Lemma 3.13. The second main component of the proof is an analysis near the critical region, more precisely in the compact set $\mathcal{X}$. We

first obtain that SGD does not spend much time in the region between $\mathcal{X}$ and $\mathcal{U}$ (i.e., in $\mathcal{X} \setminus \mathcal{U}$) either. This is achieved by the following lemma, which comes from Kifer (1988a, Lem. 5.3): given $\mathcal{U}$ an open neighborhood of $\text{crit}(f)$ and $\mathcal{X}$ a compact set containing $\mathcal{U}$, this lemma provides an exponential tail bound on the exit time from $\mathcal{X} \setminus \mathcal{U}$, which is a compact set that does not contain any critical point of f .

Lemma 3.14. *Let $\mathcal{U}$ be a neighborhood of $\text{crit } f$ and let $\mathcal{X}$ be a compact set such that $\mathcal{U} \subset \mathcal{X}$. Then there exists $a, b > 0$, such that, for η small enough, $n \geq 0$, $x \in \mathcal{K}$,*

$$\mathbb{P}_x\left(\tau_{\mathbb{R}^d \setminus \mathcal{X} \cup \mathcal{U}}^\eta > n\right) = \mathbb{P}_x\left(\sigma_{\mathcal{X} \setminus \mathcal{U}}^\eta > n\right) \leq \exp\left(-\frac{an}{\eta} + b\right). \quad (3.77)$$

or, equivalently, there are some $a, b > 0$ such that, for $\eta, \alpha > 0$ small enough, for any $x \in \mathcal{K}$,

$$\mathbb{E}_x\left[e^{\frac{\alpha \sigma_{\mathcal{X} \setminus \mathcal{U}}^\eta}{\eta}}\right] \leq e^{\frac{a\alpha}{\eta} + b}. \quad (3.78)$$

We now come to the final ingredient of the proof of Proposition 3.11: when SGD exits $\mathcal{X} \setminus \mathcal{U}$, it can go to either $\mathcal{U}$ or $\mathbb{R}^d \setminus \mathcal{X}$. We thus need to ensure that SGD goes to $\mathcal{U}$ instead of $\mathbb{R}^d \setminus \mathcal{X}$ with high probability: the following lemma provides an exponential tail bound on the probability that SGD goes to $\mathbb{R}^d \setminus \mathcal{X}$ instead of $\mathcal{U}$:

Lemma 3.15. *Let $\mathcal{U}$ be a neighborhood of $\text{crit } f$ and let $\mathcal{X}$ be a compact set such that $\mathcal{U} \subset \mathcal{X}$. There exists $\mathcal{X}'$ compact set containing $\mathcal{X}$, $\Delta > 0$, for η small enough, for any $x \in \mathcal{X}$,*

$$\mathbb{P}_x\left(\sigma_{\mathcal{X}'}^\eta < \tau_{\mathcal{U}}^\eta\right) \leq \exp\left(-\frac{\Delta}{\eta}\right). \quad (3.79)$$

The proof of Lemma 3.15 relies on several ingredients. First, Lemma 3.14 ensures that, with high probability, SGD exits $\mathcal{X} \setminus \mathcal{U}$ in a bounded number of steps, allowing us to deal with trajectories of bounded length. Then, we leverage the V_∞ (Definition 3.8) and its properties. We choose $\mathcal{X}$ large enough so that any trajectory that goes from $\mathcal{X}$ to $\mathbb{R}^d \setminus \mathcal{X}'$ without going through $\mathcal{U}$ must make V_∞ increase. (3.76) then ensures that any such trajectory must have a large action cost. The LDP estimates from Proposition 3.4(3.50b) finally guarantee that any such trajectory is unlikely, which allows us to obtain the desired exponential tail bound.

3.6.3 Estimating the transition times

We now have all the required ingredients to prove Proposition 3.11, which we restate for completeness:

Proposition 3.16. *Let $\mathcal{U}$ be a neighborhood of $\text{crit } f$ and let $\mathcal{X}$ be a compact set such that $\mathcal{U} \subset \mathcal{X}$. Then, there is some $a, b > 0$ such that, for η small enough, for any $n \geq 0$ and any $x \in \mathcal{X}$,*

$$\mathbb{P}_x\left(\tau_{\mathcal{U}}^\eta > n\right) \leq \exp\left(-\frac{an}{\eta} + b\right), \quad (3.80)$$

or, equivalently, there is some $a, b > 0$ such that, for $\eta, \alpha > 0$ small enough, for any $x \in \mathcal{X}$,

$$\mathbb{E}_x\left[e^{\frac{\alpha \tau_{\mathcal{U}}^\eta}{\eta}}\right] \leq e^{\frac{a\alpha}{\eta} + b}. \quad (3.81)$$

The proof consists in combining the three previous lemmas (Lemmas 3.13–3.15) with the strong Markov property of SGD. It goes as follows: suppose SGD starts in some $x \in \mathcal{X} \setminus \mathcal{U}$.

1. By Lemma 3.14 SGD spends a small amount of time in $\mathcal{X} \setminus \mathcal{U}$ with high probability before exiting it.
2. If SGD exists $\mathcal{X} \setminus \mathcal{U}$ through $\mathcal{U}$, we are done. Otherwise, SGD exists $\mathcal{X} \setminus \mathcal{U}$ through $\mathbb{R}^d \setminus \mathcal{X}'$ — with $\mathcal{X}'$ from Lemma 3.15 — but this happens with small probability by Lemma 3.15.
3. But, by Lemma 3.13, SGD spends only a bounded amount of time outside $\mathcal{X}$ with high probability before coming back to $\mathcal{X}$.
4. We are now back to the first step and we can iterate this reasoning until SGD exits $\mathcal{X} \setminus \mathcal{U}$ through $\mathcal{U}$.

This reasoning can be made formal by considering the moment generating function of the hitting time $\tau_{\mathcal{U}}^{\eta}$

$$\mathbb{E}_x \left[e^{\frac{\alpha \tau_{\mathcal{U}}^{\eta}}{\eta}} \right], \quad (3.82)$$

for some $\alpha > 0$ small enough. By the strong Markov property for SGD, the sketch above translates to the following (informal) recursion:

$$\begin{aligned} & \sup_{x \in \mathcal{X} \setminus \mathcal{U}} \mathbb{E}_x \left[e^{\frac{\alpha \tau_{\mathcal{U}}^{\eta}}{\eta}} \right] \\ & \leq \sup_{x \in \mathcal{X} \setminus \mathcal{U}} \left\{ \mathbb{E}_x \left[e^{\frac{\alpha \sigma_{\mathcal{X} \setminus \mathcal{U}}^{\eta}}{\eta}} \right] + \mathbb{P}_x \left(\sigma_{\mathcal{X} \setminus \mathcal{U}}^{\eta} < \tau_{\mathcal{U}}^{\eta} \right) \mathbb{E}_x \left[e^{\frac{\alpha \sigma_{\mathcal{X} \setminus \mathcal{U}}^{\eta}}{\eta}} + \mathbb{1}_{\{\sigma_{\mathcal{X} \setminus \mathcal{U}}^{\eta} < \tau_{\mathcal{U}}^{\eta}\}} \mathbb{E}_{x_{\sigma_{\mathcal{X} \setminus \mathcal{U}}^{\eta}}} \left[e^{\frac{\alpha \tau_{\mathcal{U}}^{\eta}}{\eta}} \right] \right] \right\} \\ & \leq \exp\left(\frac{a\alpha}{\eta} + b\right) + \exp\left(-\frac{\Delta}{\eta}\right) \left(\exp\left(\frac{a\alpha}{\eta} + b\right) + \sup_{z \in \mathcal{X}} \mathbb{E}_z \left[e^{\frac{\alpha \tau_{\mathcal{U}}^{\eta}}{\eta}} \right] \right). \end{aligned} \quad (3.84)$$

where we used the estimates of Lemma 3.13, Lemma 3.14, and Lemma 3.15 to bound the different terms. Making this recursion formal and solving it, we obtain the desired bound on the moment generating function of $\tau_{\mathcal{U}}^{\eta}$ of Proposition 3.11.

3.7 TRANSITION PROBABILITIES

3.7.1 Transition costs

We can now direct our attention to estimating the transition probabilities between neighborhoods of critical components: if SGD is started in some $\mathcal{C}_i$, how likely is it to reach some $\mathcal{C}_j$? We will crucially rely on the large deviations estimates obtained in § 3.3.

For this, we introduce the *transition costs* $\mathcal{B}_{i,j}$ between critical components $\mathcal{C}_i$ and $\mathcal{C}_j$. They are defined using the quasi-potential $\mathcal{B}$ (Definition 3.6) as follows:

Definition 3.9. For $i, j \in \{1, \dots, N_{\text{crit}}\}$, define the transition cost $\mathcal{B}_{i,j}$ between $\mathcal{C}_i$ and $\mathcal{C}_j$ as

$$\begin{aligned} \mathcal{B}_{i,j} &:= \inf \{ \mathcal{B}(x, x') : x \in \mathcal{K}_i, x' \in \mathcal{K}_j \} \\ &= \inf \{ \mathcal{S}_{[0,T]}(\gamma) : \gamma \in \mathcal{C}([0, T], \mathbb{R}^d), \gamma_0 \in \mathcal{C}_i, \gamma_T \in \mathcal{C}_j, T \in \mathbb{N}, T \geq 1 \} \end{aligned}$$

$$= \inf \{ \mathcal{S}_N(\xi) : \xi \in (\mathbb{R}^d)^N, \xi_0 \in \mathcal{C}_i, \xi_{N-1} \in \mathcal{C}_j, N \geq 1 \}. \quad (3.85)$$



Figure 12: Illustration of a path that goes from $\mathcal{C}_i$ to $\mathcal{C}_j$.

The transition costs $(\mathcal{B}_{i,j})_{i,j}$ characterize the transition probabilities of the induced chain $(z_n)_{n \geq 0}$ (Definition 3.5) between neighborhoods of critical components and, more precisely, the N -step transition probabilities of the $(z_n)_{n \geq 0}$, denoted by $\mathbb{Q}^N(x, \cdot)$: for N large enough, for $x \in \mathcal{C}_i$,

$$\mathbb{Q}^N(x, \mathcal{C}_j^\delta) = \mathbb{P}_x(x_N^\eta \in \mathcal{C}_j^\delta) \approx \exp\left(-\frac{\mathcal{B}_{i,j}}{\eta}\right). \quad (3.86)$$

Therefore, the transition cost $\mathcal{B}_{i,j}$ can be interpreted as the cost for SGD to go from $\mathcal{C}_i$ to $\mathcal{C}_j$ in a large number of steps: the higher $\mathcal{B}_{i,j}$, the less likely it is for SGD to go from $\mathcal{C}_i$ to $\mathcal{C}_j$ in a large number of steps.

Again, $\mathcal{B}_{i,j}$ inherits the properties of $\mathcal{S}$ that are detailed in §3.3.1. In particular, if there exists a solution of the gradient flow that goes from $\mathcal{C}_i$ to $\mathcal{C}_j$, then $\mathcal{B}_{i,j} = 0$ and SGD going from $\mathcal{C}_i$ to $\mathcal{C}_j$ is very likely. Conversely, in the case of Gaussian noise §3.3.2, (3.38) implies that $\mathcal{B}_{i,j}$ is lower-bounded as:

$$\mathcal{B}(x, x') \geq \frac{2(f(\mathcal{C}_i) - f(\mathcal{C}_j))}{\sigma^2}. \quad (3.87)$$

Hence, if $\mathcal{C}_i$ sits at the bottom of a hill and $\mathcal{C}_j$ sits at the top of that hill, then $\mathcal{B}_{i,j}$ is large.

A priori, nothing guarantees the transition costs $\mathcal{B}_{i,j}$ will be finite. For simplicity, we will assume throughout that they are.

Assumption 3.10. For any $i \in \{1, \dots, K\}$ and $j \in \{1, \dots, N_{\text{crit}}\}$, $\mathcal{B}_{i,j} < +\infty$.

In light of (3.86), Assumption 3.10 ensures that it is possible for SGD to go from any critical component to any other critical component with non-negligible zero (though it can still be very small). This is a technical assumption and, it can be guaranteed by requiring that for each pair of components $(\mathcal{C}_i, \mathcal{C}_j)$, there exists a path γ from $\mathcal{C}_i$ to $\mathcal{C}_j$ along which the noise can actively steer SGD (see Appendix B.2.4 for details). Also note that if Assumption 3.10 does not hold, one could still carry out the analysis by restricting attention to the communicating classes of the transition graph, applying our results within each class separately; we do not pursue this here for simplicity.

3.7.2 Transition probabilities

We here explain how to obtain a formal version of (3.86). For this, we adapt to our context Kifer (1988a, Lem. 5.4) and simplify it using ideas from Freidlin and Wentzell (1998b, Chap. 6).

We introduce the auxiliary transition costs $\tilde{\mathcal{B}}_{i,j}$: it is the analogue of $\mathcal{B}_{i,j}$ when we only consider paths that do not go through any other critical component than $\mathcal{C}_i$ and $\mathcal{C}_j$.

Definition 3.11 (Freidlin and Wentzell (1998b, Chap. 6,§2)). For i, j ,

$$\tilde{\mathcal{B}}_{i,j} := \inf \left\{ S_N(\xi) : N \geq 1, \xi \in (\mathbb{R}^d)^N, \xi_0 \in \mathcal{C}_i, \xi_{N-1} \in \mathcal{C}_j, \xi_n \notin \bigcup_{l \neq i,j} \mathcal{C}_l \text{ for } n = 1, \dots, N-2 \right\}. \quad (3.88)$$

The distinction between $\mathcal{B}_{i,j}$ and $\tilde{\mathcal{B}}_{i,j}$ is subtle but important. $\mathcal{B}_{i,j}$ is the minimum cost to go from $\mathcal{C}_i$ to $\mathcal{C}_j$ along any path, including those that pass through other critical components $\mathcal{C}_l$ along the way. By contrast, $\tilde{\mathcal{B}}_{i,j}$ is the minimum cost subject to the constraint that the path avoids all other critical components: it measures the cost of a *direct* transition from $\mathcal{C}_i$ to $\mathcal{C}_j$. In particular, $\mathcal{B}_{i,j} \leq \tilde{\mathcal{B}}_{i,j}$ since $\mathcal{B}_{i,j}$ optimizes over a strictly larger set of paths.

Lemma 3.17 shows that $\mathcal{B}_{i,j}$ equal the shortest-path distance from $\mathcal{C}_i$ to $\mathcal{C}_j$ in the weighte graph whose edge weights are the direct transition costs $\tilde{\mathcal{B}}_{l,l'}$: it is the cost of the cheapest *sequence of direct transitions* connecting $\mathcal{C}_i$ to $\mathcal{C}_j$.

Lemma 3.17. For any $i, j \in I$,

$$\begin{aligned} \mathcal{B}_{i,j} &= \min \left\{ \sum_{l=0}^{n-2} \tilde{\mathcal{B}}_{i_l, i_{l+1}} : i_0 = i, i_{n-1} = j, i_l \in I \text{ for } l = 1, \dots, n-2, n \geq 1 \right\} \\ &= \min \left\{ \sum_{l=0}^{N_{\text{crit}}-2} \tilde{\mathcal{B}}_{i_l, i_{l+1}} : i_0 = i, i_{N_{\text{crit}}-1} = j, i_l \in I \text{ for } l = 1, \dots, N_{\text{crit}}-2 \right\}. \end{aligned} \quad (3.89)$$

The auxiliary costs $\tilde{\mathcal{B}}_{i,j}$ are the natural objects for estimating one-step transition probabilities of the induced chain: since the chain is stopped upon entering any neighborhood $\mathcal{C}_l^\delta$, a single transition corresponds precisely to a path that does not pass through any other component. **Lemma 3.17** then recovers $\mathcal{B}_{i,j}$ from these one-step costs via the shortest-path formula above. We now make this quantitative: we first estimate the one-step transition probabilities in terms of $\tilde{\mathcal{B}}_{i,j}$ (**Lemmas 3.18** and **3.19**), then combine them via **Lemma 3.17** to obtain estimates in terms of $\mathcal{B}_{i,j}$ (**Proposition 3.20**).

Lemma 3.18. Fix $\varepsilon > 0$. Let $\eta, \delta > 0$ be small enough. Then, for all $i, j \in \{1, \dots, N_{\text{crit}}\}$,

$$\forall x \in \mathcal{C}_i^\delta, \quad \mathbb{Q}(x, \mathcal{C}_j^\delta) \leq \exp \left(-\frac{\tilde{\mathcal{B}}_{i,j}}{\eta} + \frac{\varepsilon}{\eta} \right). \quad (3.90)$$

Proof sketch. The proof of this upper bound closely follows **Kifer (1988a)**, though adapting his arguments to our setting requires all the tools developed so far. It consists in bounding the probability that $z_1 \in \mathcal{C}_j^\delta$ given that $z_0 = x \in \mathcal{C}_i^\delta$, and combines on three key ingredients:

- By **Lemma 3.10**, any trajectory from $\mathcal{C}_i^\delta$ to $\mathcal{C}_j^\delta$ can be modified, at negligible cost, into one that starts in $\mathcal{C}_i$ and ends in $\mathcal{C}_j$. By definition of $\tilde{\mathcal{B}}_{i,j}$ (**Definition 3.11**), any such trajectory therefore has action cost at least $\tilde{\mathcal{B}}_{i,j}$.
- By **Proposition 3.11**, with high probability, any trajectory going from $\mathcal{C}_i^\delta$ to $\mathcal{C}_j^\delta$ does so in a bounded number of steps N .
- The LDP upper bound ((3.50b) in **Proposition 3.4**) then yields that the probability of going from $\mathcal{C}_i^\delta$ to $\mathcal{C}_j^\delta$ in N steps is at most $\exp(-(\tilde{\mathcal{B}}_{i,j} - \varepsilon)/\eta)$.

■

We now provide a lower bound on the transition probabilities $\mathbb{Q}^N(x, \mathcal{C}_j^\delta)$ for N large enough in terms of the auxiliary transition costs $\tilde{\mathcal{B}}_{i,j}$.

Lemma 3.19. *With the setting of Lemma 3.18, in addition, there exists N such that,*

$$\forall x \in \mathcal{C}_i^\delta, \quad \mathbb{Q}^N(x, \mathcal{C}_j^\delta) \geq \exp\left(-\frac{\tilde{\mathcal{B}}_{i,j}}{\eta} - \frac{\varepsilon}{\eta}\right). \quad (3.91)$$

Proof sketch. The proof of this lower bound closely follows Kifer (1988a), though here again the adaptation to our setting makes essential use of all the tools developed so far. It consists in explicitly constructing a trajectory from $x \in \mathcal{C}_i^\delta$ to $\mathcal{C}_j^\delta$ in N steps with action cost approximately $\tilde{\mathcal{B}}_{i,j}$, and combines again three key ingredients:

- By definition of $\tilde{\mathcal{B}}_{i,j}$ (Definition 3.11), there exists a trajectory from some point $\xi_0 \in \mathcal{C}_i$ to $\mathcal{C}_j$ with action cost approximately $\tilde{\mathcal{B}}_{i,j}$ that avoids all other critical components.
- By Lemma 3.10 and Lemma 3.9, one can prepend a trajectory of bounded length and negligible cost from $x \in \mathcal{C}_i^\delta$ to $\xi_0 \in \mathcal{C}_i$ that remains close to $\mathcal{C}_i$. The concatenation yields a trajectory from x to $\mathcal{C}_j$ with action cost approximately $\tilde{\mathcal{B}}_{i,j}$ that avoids all other critical components.
- The LDP lower bound ((3.50a) in Proposition 3.4) then yields that the probability of following such a trajectory is at least $\exp(-(\tilde{\mathcal{B}}_{i,j} + \varepsilon)/\eta)$.

■

Proposition 3.20 is the main result of this section: it exactly characterizes the N -transition probabilities of the induced chain $(z_n)_{n \geq 0}$ between neighborhoods of critical components in terms of the transition costs $\mathcal{B}_{i,j}$:

Proposition 3.20. *With the setting of Lemma 3.18, in addition, there exists N such that,*

$$\forall x \in \mathcal{C}_i^\delta, \quad \exp\left(-\frac{\mathcal{B}_{i,j} + \varepsilon}{\eta}\right) \leq \mathbb{Q}^N(x, \mathcal{C}_j^\delta) \leq \exp\left(-\frac{\mathcal{B}_{i,j} - \varepsilon}{\eta}\right). \quad (3.92)$$

To obtain Proposition 3.20, we combine Lemma 3.17 with iterated applications of Lemmas 3.18 and 3.19.

3.8 SUMMARY

The goal of this chapter was to develop the proof infrastructure for the main results on SGD by adapting the Freidlin–Wentzell–Kifer program of Freidlin and Wentzell (1998b), Kifer (1988a) to the SGD setting. The central idea is to reduce the analysis of the long-run behavior of SGD to that of a finite-state Markov chain — the *induced chain* — whose transition probabilities are governed by large deviation costs between the connected components of $\text{crit } f$.

The first step is a sample-path large deviation principle for the accelerated SGD process (§ 3.3), with action functional

$$\mathcal{S}_T(\gamma) = \int_0^T \mathcal{L}(\gamma(t), \dot{\gamma}(t)) \, dt \quad (3.93)$$

built from the Hamiltonian of the noise. Armed with this LDP, we carry out the approach of Freidlin–Wentzell:

- Step 1: Reduction to a finite-state space (§ 3.5) The behavior of SGD depends only on which component the iterates are near, not on the precise location within it (Lemma 3.9 and Lemma 3.10). This allows us to treat the induced chain as a Markov chain on the finite set of critical components $\{\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}\}$ and to define transition costs between them independently of the neighborhoods chosen.
- Step 2: Transition times (§ 3.6) SGD returns to any neighborhood of crit f with exponential tails on the return time (Proposition 3.11), showing that the induced chain faithfully represents the long-run dynamics.
- Step 3: Transition probabilities (§ 3.7) The transition probabilities of the induced chain satisfy the two-sided estimate (Proposition 3.20): for N large enough, for $x \in \mathcal{C}_i^\delta$,

$$\exp\left(-\frac{\mathcal{B}_{i,j} + \varepsilon}{\eta}\right) \leq \mathbb{Q}^N(x, \mathcal{C}_j^\delta) \leq \exp\left(-\frac{\mathcal{B}_{i,j} - \varepsilon}{\eta}\right) \quad (3.94)$$

where $\mathcal{B}_{ij}$ is the minimum cost, as measured by the action functional, to travel from $\mathcal{K}_i$ to $\mathcal{K}_j$.

Extending the Freidlin–Wentzell approach to SGD requires addressing three main difficulties: the unboundedness of the domain, the boundedness and non-Gaussianity of the noise, and the discrete-time nature of the process. Together, the results of this chapter are the key ingredients for the proofs of the main results on the invariant measure and the hitting-time asymptotics, presented in Chapter 4 and Chapter 5.

THE INVARIANT MEASURE OF SGD

4.1	Characterizing the invariant measure	51
4.1.1	From induced chain to SGD	51
4.1.2	Relative probabilities of critical components	55
4.1.3	Global concentration near minimizing components	57
4.2	Applications and implications	59
4.2.1	SGD with Gaussian noise	59
4.2.2	Local dependencies	64
4.2.3	Convergence of empirical measures	65
4.2.4	Dissipative systems	66

In this chapter, we present our main results on the invariant measure of (SGD). The material is taken from [Azizian et al. \(2024c\)](#) but completely rewritten; in particular, we highlight how the results follow directly from the essential ingredients of [Chapter 3](#).

In [§ 4.1](#), we derive the main results on the invariant measure. We begin, in [§ 4.1.1](#), by establishing the key formula relating the invariant measure of SGD to that of its induced chain. In [§ 4.1.2](#), we apply this formula to characterize the relative probabilities of the different critical components. In [§ 4.1.3](#), we show that the iterates are exponentially unlikely to be observed far from $\text{crit } f$ and thus from the ground states. In [§ 4.2](#), we examine the consequences and applications of these results. We first compute the energy levels explicitly for SGD with Gaussian noise ([§ 4.2.1](#)), then study how the local geometry of the noise controls the energy levels ([§ 4.2.2](#)). We then extend all four main results to the mean occupation measures ([§ 4.2.3](#)). Finally, we show how our theory connects to some deep learning phenomena ([§ 4.2.4](#)) under an additional dissipativity assumption.

4.1 CHARACTERIZING THE INVARIANT MEASURE

4.1.1 From induced chain to SGD

To motivate the main formula for the invariant measure of the induced chain, we begin by considering an actual finite-state Markov chain and present a combinatorial representation of invariant measures for finite-state Markov chains. This formula can be seen as a generalization of Kirchhoff's matrix tree theorem. Its statement for the invariant distribution of Markov chains is from ([Hill, 1966](#), [Leighton and Rivest, 1986](#)); see also [Freidlin and Wentzell \(1998b, Chap. 6, Lem. 3.1\)](#) for a more general version. We present it here in a self-contained way for the reader's convenience.

Consider a Markov chain on the state space $I = \{1, \dots, N_{\text{crit}}\}$ with transition matrix $(p_{ij})_{i,j \in I}$. An invariant probability measure $\mu = (\mu_i)_{i \in I}$ satisfies, for every $i \in I$,

$$\mu_i = \sum_{j \in I} \mu_j p_{ji}. \quad (4.1)$$

Equivalently, using $\sum_{j \in I} p_{ij} = 1$, this can be rewritten in the global balance form: for every $i \in I$,

$$\sum_{j \neq i} p_{ij} \mu_i = \sum_{j \neq i} p_{ji} \mu_j. \quad (4.2)$$

These balance equations form a linear system for the unknown masses $(\mu_i)_{i \in I}$ and characterize the invariant measure implicitly. Indeed, (4.1) can be written as

$$\mu(I - P) = 0, \quad \sum_{i \in I} \mu_i = 1.$$

These equations could be solved exactly and their solution may be expressed through minors or cofactors. The resulting formulas are exact, but not especially explicit.

For our purposes, we need an explicit form, one that reveals how the whole transition structure contributes to the mass at a given state. Fortunately, it is possible to obtain a combinatorial expression for the same solution (Hill, 1966, Leighton and Rivest, 1986). It rewrites the stationary mass at i as the total weight of all directed spanning trees rooted at i , that is, of all global flow structures that funnel the chain toward i .

This formula involves *directed spanning trees rooted at a node i* : these are directed graphs on I in which every vertex $j \neq i$ has exactly one outgoing edge, i has no outgoing edge, and there is no directed cycle. We denote by $\mathcal{T}_i$ the set of directed spanning trees rooted at i . For each $i \in I$, define the unnormalized weight

$$\pi_i := \sum_{t \in \mathcal{T}_i} \prod_{(j \rightarrow l) \in t} p_{jl}, \quad (4.3)$$

Each tree may be interpreted as a minimal global routing pattern by which every state eventually feeds into i ; its weight is the product of the transition probabilities along its edges. Summing over all such trees therefore measures the total strength of all coherent ways in which the dynamics can direct mass toward i . Setting

$$\mu_i := \frac{\pi_i}{\sum_{j \in I} \pi_j},$$

yields an invariant probability distribution for the chain. For completeness, we provide a self-contained proof of this fact, following the argument of Cerf (2019).

Proof the claim. To check that μ is an invariant distribution, it is enough to show that the unnormalized weights satisfy the global balance equations: for all $i \in I$,

$$\sum_{j \neq i} p_{ij} \pi_i = \sum_{j \neq i} p_{ji} \pi_j. \quad (4.4)$$

To verify (4.4), we show that both sides are equal to

$$\sum_g \prod_{(j \rightarrow l) \in g} p_{jl}, \quad (4.5)$$

where the sum ranges over all directed graphs g on I in which every vertex has exactly one outgoing edge¹ and the unique directed cycle passes through i .

For the left-hand side, each pair $(j \neq i, t \in \mathcal{T}_i)$ produces such a graph by adding the edge $i \rightarrow j$ to t . This creates exactly one directed cycle, which necessarily passes through i , and every vertex still has exactly one outgoing edge. Conversely, every graph g contributing to (4.5) arises uniquely in this way: removing the outgoing edge of i breaks the unique cycle and yields a directed spanning tree rooted at i . Therefore, the left-hand side of (4.4) is equal to (4.5).

The same argument applies to the right-hand side. Each pair $(j \neq i, t \in \mathcal{T}_j)$ produces such a graph by adding the edge $j \rightarrow i$ to t . Again, this creates a unique directed cycle passing through i , and every graph counted in (4.5) is obtained in a unique way by removing the edge entering i along the cycle. Hence the right-hand side of (4.4) is also equal to (4.5).

This proves (4.4), and therefore the normalized weights define an invariant distribution. ■

In the preceding chapter, we reduced the long-run behavior of (SGD) to the transitions of its iterates between the components $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ of $\text{crit } f$, yielding an induced Markov chain $(z_n)_n$ that is very close to the finite-state setting above. By Proposition 3.20, its transition probabilities satisfy, for all $i \neq j$,

$$p_{ij} \sim \exp\left(-\frac{\mathcal{B}_{i,j}}{\eta}\right). \quad (4.6)$$

Applying the computations above², substituting into (4.3), the unnormalized weight of $\mathcal{K}_i$ should satisfy

$$\begin{aligned} \pi_i &\sim \sum_{t \in \mathcal{T}_i} \prod_{(j \rightarrow l) \in t} \exp\left(-\frac{\mathcal{B}_{j,l}}{\eta}\right) \\ &= \sum_{t \in \mathcal{T}_i} \exp\left(-\frac{1}{\eta} \sum_{(j \rightarrow l) \in t} \mathcal{B}_{j,l}\right) \\ &\sim \exp\left(-\frac{1}{\eta} \min_{t \in \mathcal{T}_i} \sum_{(j \rightarrow l) \in t} \mathcal{B}_{j,l}\right), \end{aligned} \quad (4.7)$$

where the last step uses the fact that the sum concentrates on the minimum-weight tree for small enough η . This minimum total tree weight is precisely the *energy* of component $\mathcal{K}_i$. It is the quantity that determines the weight of $\mathcal{C}_i$ in the invariant measure of SGD.

In the rest of this section, we introduce the necessary tools to make the above argument (4.7) rigorous. First, the transition costs $\mathcal{B}_{i,j}$ can be organized into a directed graph, which we call the *transition graph* of (SGD).

Definition 4.1. The *transition graph* of (SGD) is the complete weighted directed graph $g = (\mathcal{V}, \mathcal{E})$ defined as follows:

- 1 A directed graph with exactly one outgoing edge for each node and no self-loop must have exactly one directed cycle.
- 2 Note that these computations do not formally apply: $(z_n)_n$ is not formally a finite-state Markov chain, the results of Chapter 3 only show that it behaves like one up to some multiplicative error terms. Freidlin and Wentzell (1998b, Chap. 6, Lem. 3.1) provides a precise way to make these approximations rigorous, see Proposition 4.1 below.

- (a) The vertex set is $\mathcal{V} = I$, with one vertex per component of crit f .
- (b) The edge set is $\mathcal{E} = \{(i, j) : i, j \in I, i \neq j\}$, with one directed edge per ordered pair of distinct components.
- (c) The weight of each directed edge $(i \rightarrow j)$ is the transition cost $\mathcal{B}_{i,j}$.

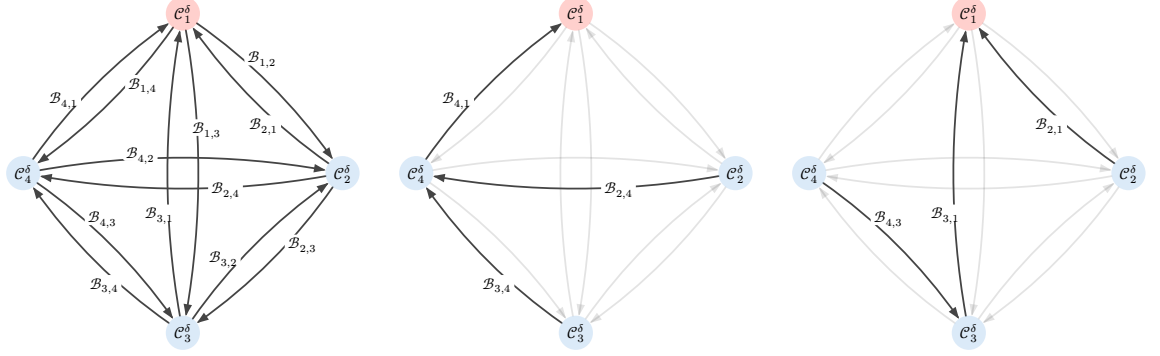


Figure 13: Illustration of the transition graph on $K = 4$ components. *Left:* the complete transition graph, with one directed edge per ordered pair (i, j) carrying weight $\mathcal{B}_{i,j}$. *Center and right:* directed spanning trees $t \in \mathcal{T}_i$ rooted at 1 (highlighted in red), with every other vertex having exactly one outgoing edge and all paths leading to i .

We can now define the energy of a component C_i , which appears in (4.7).

Definition 4.2 ((Freidlin and Wentzell, 1998b, Chap. 6, §4)). Define, for every $i \in I$,

$$E_i = E(C_i) := \min_{t \in \mathcal{T}_i} \sum_{(j \rightarrow i) \in t} \mathcal{B}_{j,i}, \quad (4.8)$$

where $\mathcal{T}_i$ denotes the set of directed trees³ rooted at i in the complete graph on I .

See Figure 13 for an illustration of the transition graph and of the directed spanning trees rooted at a given vertex.

To simplify the statements, for the rest of this chapter, we fix a precision level $\varepsilon > 0$, and consider sufficiently small C_i^δ δ -neighborhoods of C_i , $i = 1, \dots, N_{\text{crit}}$ and sufficiently small $\eta > 0$.

The following proposition makes the approximations precise. It is a consequence of Freidlin and Wentzell (1998b, Chap. 6, Lem. 3.1), which provides explicit bounds on the invariant measure of a Markov chain whose state space is partitioned into finitely many components, given corresponding bounds on the transition probabilities between components. We apply this to the induced chain $(z_n)_n$ on $\mathcal{C}^\delta = \bigcup_{i=1}^K C_i^\delta$, using the transition probability estimates of Proposition 3.20.

This proposition thus fully characterizes ν_∞^η (which is supported on $\mathcal{C}^\delta = \bigcup_{i=1}^K C_i^\delta$).

Proposition 4.1. Consider ν_∞^η an invariant probability measure for $(z_n)_n$ the induced chain on $\mathcal{C}^\delta$. For any $i \in I$,

$$\exp\left(-\frac{E_i - \min_{j \in I} E_j + \varepsilon}{\eta}\right) \leq \nu_\infty^\eta(C_i^\delta) \leq \exp\left(-\frac{E_i - \min_{j \in I} E_j - \varepsilon}{\eta}\right).$$

³ A directed tree rooted at i is graph without cycles and such that every node other than i has exactly one outgoing edge, and i has no outgoing edge.

It now remains to relate ν_∞^η to the invariant measures of (SGD) itself. This is provided by the following key lemma, see e.g., (Douc et al., 2018, Thm. 3.6.5). It also corresponds to Kifer (1990, Cor. 3.1) adapted to our non-compact setting. It is also linked to results from Khasminskii (2012, Thm. 4.1) which studies the continuous-time case and from Meyn and Tweedie (1993, Thm. 10.4.9) which studies the irreducible case.

Lemma 4.2. *Consider μ_∞^η an invariant probability measure for (SGD). Then, ν_∞^η , the restriction of $\mu_\infty^\eta / \mu_\infty^\eta(\mathcal{C}^\delta)$ to $\mathcal{C}^\delta$ is an invariant measure for the induced chain on $\mathcal{C}^\delta$. Moreover, for any measurable set $H \subset \mathbb{R}^d$,*

$$\frac{\mu_\infty^\eta(H)}{\mu_\infty^\eta(\mathcal{C}^\delta)} = \int_{\mathcal{C}^\delta} \mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in H\} \right] d\nu_\infty^\eta(x). \quad (4.9)$$

This lemma states that μ_∞^η invariant measures of SGD roughly correspond to the invariant measures of the induced chain ν_∞^η provided that the chain does not spend too much time outside of $\mathcal{C}^\delta$ so that the expectation

$$\mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in H\} \right] \quad (4.10)$$

can be adequately controlled. Proposition 3.11, which bounds the time (SGD) spends outside of $\mathcal{C}^\delta$, is key in that.

4.1.2 Relative probabilities of critical components

We can now present our main results. We begin by estimating the relative probabilities of the different components $\mathcal{C}_i$ for $i = 1, \dots, K$. For this, we apply Lemma 4.2 with $H \leftarrow \mathcal{C}_i^\delta$, δ -neighborhood of $\mathcal{C}_i$, to obtain, up to the normalization constant,

$$\mu_\infty^\eta(\mathcal{C}_i^\delta) \propto \int_{\mathcal{C}^\delta} \mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in \mathcal{C}_i^\delta\} \right] d\nu_\infty^\eta(x). \quad (4.11)$$

But $\mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in \mathcal{C}_i^\delta\} \right]$ is lower-bounded by 1 by definition and uniformly upper-bounded by $e^{\frac{\varepsilon}{\eta}}$ thanks to Proposition 3.11. Hence, $\mu_\infty^\eta(\mathcal{C}_i^\delta)$ coincides with $\nu_\infty^\eta(\mathcal{C}_i^\delta)$ up to multiplicative factors. Combined with Proposition 4.1, this gives the following main result:

Theorem 4.3. *Let μ_∞^η be an invariant measure under (SGD), For any $i \in I$,*

$$\exp\left(-\frac{E_i - \min_{j \in I} E_j}{\eta} - \frac{\varepsilon}{\eta}\right) \leq \mu_\infty^\eta(\mathcal{U}_i) \leq \exp\left(-\frac{E_i - \min_{j \in I} E_j}{\eta} + \frac{\varepsilon}{\eta}\right). \quad (4.12)$$

More compactly, with notation as above, we have:

$$\mu_\infty^\eta(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right). \quad (4.13)$$

Theorem 4.3 states that the long-run distribution of (SGD) around the components of crit f is, to exponential order, a Boltzmann–Gibbs measure with energy levels (E_i) at temperature η (Landau and Lifshitz, 1976). Its content is perhaps most clearly expressed as a statement about *relative* probabilities: for any two components $\mathcal{K}_i$ and $\mathcal{K}_j$, up to multiplicative error,

$$\frac{\mu_\infty^\eta(\mathcal{U}_i)}{\mu_\infty^\eta(\mathcal{U}_j)} \sim \exp\left(-\frac{E_i - E_j}{\eta}\right), \quad (4.14)$$

so the ratio of time spent near $\mathcal{K}_i$ versus $\mathcal{K}_j$ is determined, to leading exponential order, entirely by the difference $E_i - E_j$. The energy levels $(E_i)_i$ are not (necessarily) the objective values (f_i) : they arise from a variational problem over the transition graph, where E_i is the minimum total cost of routing the entire chain toward $\mathcal{K}_i$, as measured by the quasi-potentials $(\mathcal{B}_{i,j})$. Since these quasi-potentials encode both the geometry of f and the statistics of the noise, the ground state — the minimizer of E_i — depends on the noise model and need not coincide with the global minimizer of f (see §§ 4.2.1 and 4.2.2 for a detailed discussion of this point).

We note that **Theorem 4.3** applies to *any* invariant measure of (SGD): our assumptions guarantee existence (see **Lemma A.3.19** in **Appendix A**) but not uniqueness, so the theorem characterizes all of them simultaneously. Regarding precision, the tolerance ε enters (4.14) as a multiplicative factor $e^{\pm 2\varepsilon/\eta}$; more explicitly,

$$e^{-(E_i - E_j + 2\varepsilon)/\eta} \leq \frac{\mu_\infty^\eta(\mathcal{U}_i)}{\mu_\infty^\eta(\mathcal{U}_j)} \leq e^{-(E_i - E_j - 2\varepsilon)/\eta}, \quad (4.15)$$

which is negligible relative to the leading term whenever $E_i - E_j$ is positive. The theorem holds for any noise model satisfying our standing assumptions, with all model-specific information absorbed into the quasi-potentials $(\mathcal{B}_{i,j})$: the Boltzmann–Gibbs form is the same regardless of the noise distribution; only the numerical values of $(E_i)_i$ change.

A natural question is whether the non-minimizing components of f — saddle points and local maxima — are selected against under μ_∞^η . To make this precise, we recall the following definitions.

Definition 4.3. A component $\mathcal{K}$ of crit f is called *minimizing* if there exists a neighborhood $\mathcal{U}$ of $\mathcal{K}$ such that $\mathcal{K} = \arg \min_{x \in \mathcal{U}} f(x)$; it is called *non-minimizing* otherwise.

A component $\mathcal{K}$ is thus minimizing if it is a (strict) local minimum of f ; saddles and local maxima are non-minimizing. This notion is closely linked to the notion of attracting components under the gradient flow of f :

Definition 4.4. A component $\mathcal{K}$ of crit f is called *attracting* if there exists a neighborhood $\mathcal{U}$ of $\mathcal{K}$ such that, for any $x \in \mathcal{U}$, the gradient flow $(\Theta_t(x))_{t \geq 0}$ converges to $\mathcal{K}$ as $t \rightarrow +\infty$.

In our setting, these two notions coincide: $\mathcal{K}$ is minimizing if and only if it is attracting. The key fact underpinning this equivalence is that, in our setting, the gradient flow of f converges to crit f from any initial point, and, by definition, the components of crit f are separated. Therefore, for any initial point x , the gradient flow converges to one of the components $\mathcal{C}_1, \dots, \mathcal{C}_K$ of crit f : for any $x \in \mathbb{R}^d$, there exists $i \in I$ such that

$$d(\Theta_t(x), \mathcal{C}_i) \rightarrow 0 \quad \text{as } t \rightarrow +\infty. \quad (4.16)$$

A non-minimizing component therefore cannot attract all nearby trajectories: nearby points with strictly lower objective value must eventually flow to another component.

This equivalence has a crucial consequence: it also implies that $\mathcal{C}_i$ is attracting / minimizing if and only if

$$\mathcal{B}_{i,j} > 0 \quad \text{for all } j \neq i. \quad (4.17)$$

Informally: if $\mathcal{C}_i$ is not attracting, there exists a gradient-flow trajectory that starts arbitrarily close to $\mathcal{C}_i$, has decreasing objective value, and converges to some other component $\mathcal{C}_j$; since the quasi-potential $\mathcal{B}_{i,j}$ measures the cost of transitioning from $\mathcal{C}_i$ to $\mathcal{C}_j$, this forces $\mathcal{B}_{i,j} = 0$. Iterating this argument, for any non-attracting $\mathcal{C}_i$ there exists an attracting $\mathcal{C}_j$ with $\mathcal{B}_{i,j} = 0$. Now, since $\mathcal{C}_j$ is attracting, we have that $\mathcal{B}_{j,l} > 0$ for all $l \neq j$. Thus, any spanning tree rooted at $\mathcal{C}_i$ must include an edge leaving $\mathcal{C}_j$, incurring cost at least $\min_{l \neq j} \mathcal{B}_{j,l} > 0$; by contrast, any spanning tree rooted at $\mathcal{C}_j$ can route $\mathcal{C}_i$ toward $\mathcal{C}_j$ at zero cost via $\mathcal{B}_{i,j} = 0$. Hence, it holds that

$$E_i > E_j. \quad (4.18)$$

Together with [Theorem 4.3](#), this immediately implies that non-minimizing components are exponentially suppressed under μ_∞^η :

Theorem 4.4. *Suppose that μ_∞^η is invariant under (SGD), and let $\mathcal{C}_i$ be a non-minimizing component of f . Then, there exists a minimizing component $\mathcal{C}_j$ of f and a positive constant⁴ $c > 0$ such that*

$$\frac{\mu_\infty^\eta(\mathcal{C}_i^\delta)}{\mu_\infty^\eta(\mathcal{C}_j^\delta)} \leq \exp\left(-\frac{c + \varepsilon}{\eta}\right) \quad (4.19)$$

In particular, in the limit $\eta \rightarrow 0$, we have $\mu_\infty^\eta(\mathcal{U}_i) \rightarrow 0$.

This avoidance principle is particularly important because it shows that (SGD) is far less likely to be observed near a non-minimizing components of crit f relative to a minimizing one. In this regard, [Theorem 4.4](#) complements a broad range of avoidance results in the literature [Pemantle \(1990\)](#), [Ge et al. \(2015a\)](#), [Jin et al. \(2017\)](#), [Mertikopoulos et al. \(2020\)](#), [Hsieh et al. \(2021, 2023\)](#) without requiring any of the “strict saddle” assumptions that are standard in this context.

That being said, [Theorem 4.3](#) and [Theorem 4.4](#) leave open the possibility that the energy landscape of (SGD) contains *non-critical* low-energy regions that nonetheless get a significant amount of probability under (SGD); put differently, [Theorem 4.3](#) and [Theorem 4.4](#) do not rule out the eventuality that, in the long run, x_n may still be observed with non-vanishing probability far from the critical region of f . This is addressed in the next section.

4.1.3 Global concentration near minimizing components

In this section, we show that the iterates of (SGD) are exponentially unlikely to be observed far from crit f and as a consequence, the results of the previous section fully characterize the long-run distribution of (SGD).

We begin by showing that invariant measures are exponentially concentrated around crit f . [Lemma 4.2](#) reduces this to bounding $\mathbb{E}_x[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in D\}]$

⁴ Independent of $\varepsilon, \eta, \delta$.

for D far from crit f . The key tool is the following lemma, which bounds the probability that the iterates reach a set D before returning to the critical neighborhood.

Lemma 4.5. *For any $i \in I$, $x \in \mathcal{C}_i^\delta$*

$$\mathbb{P}_x\left(\tau_D^\eta < \tau_{\mathcal{C}^\delta}^\eta\right) \leq \exp\left(-\frac{\mathcal{B}(\mathcal{C}_i, \mathcal{U}_\delta(D)) - \varepsilon}{\eta}\right), \quad (4.20)$$

where

$$\mathcal{B}(\mathcal{C}_i, \mathcal{U}_\delta(D)) := \inf_{x \in \mathcal{C}_i, x' \in \mathcal{U}_\delta(D)} \mathcal{B}_{i,j}(x, x') \quad (4.21)$$

This lemma is natural in view of the LDP (Proposition 3.4) and the definition of the quasi-potential (Definition 3.6), which suggests that the probability of reaching D from $\mathcal{C}_i$ should be approximately $\exp\left(-\frac{\mathcal{B}(\mathcal{C}_i, D)}{\eta}\right)$. Again, the proof of this lemma also relies on Proposition 3.11 to be able to apply the LDP on a finite time horizon.

In the formula from Lemma 4.2, if the chain returns to $\mathcal{C}^\delta$ without ever visiting D , the contribution $\sum \mathbb{1}\{x_n \in D\}$ is zero. So Lemma 4.5 bounds the probability of a non-zero contribution, and combined with Proposition 3.11 which controls the total excursion length, this makes $\mu_\infty^\eta(D)$ exponentially small whenever $\mathcal{B}(\mathcal{K}_i, D) > 0$ for all i . For minimizing components, similar to the earlier discussion, the quasi-potential to any set outside a neighborhood is positive: if $\mathcal{K}_i$ is minimizing, then $\mathcal{B}(\mathcal{K}_i, \mathbb{R}^d \setminus \mathcal{U}) > 0$ for any sufficiently small neighborhood $\mathcal{U}$ of $\mathcal{K}_i$. Intuitively, leaving a local minimum requires increasing the loss value which cannot be done with zero action cost.

Thus, leveraging Lemma 4.5 with D as the whole space except a neighborhood of the minimizing components, one obtains:

Theorem 4.6. *Let μ_∞^η be an invariant measure for (SGD). Then, there exists a constant⁵ $c > 0$ such that*

$$\mu_\infty^\eta(\mathcal{C}_{\min}^\delta) \geq 1 - e^{-c/\eta}, \quad \text{where } \mathcal{C}_{\min}^\delta := \bigcup_{i: \mathcal{K}_i \text{ minimizing}} \mathcal{C}_i^\delta \quad (4.22)$$

and, therefore,

$$\mu_\infty^\eta(\mathcal{C}^\delta) \geq 1 - e^{-c/\eta}, \quad \text{where } \mathcal{C}^\delta = \bigcup_{i=1}^K \mathcal{C}_i^\delta. \quad (4.23)$$

This theorem thus shows that, in the long run, the iterates of (SGD) are exponentially more likely to be observed in the vicinity of crit f rather than far from it, and exponentially more likely to be observed near a minimum of f rather than a saddle-point (or a local maximizer).

Our next result can be seen as joint consequence of Theorem 4.3 and Theorem 4.6 as it shows that the long-run distribution of (SGD) is exponentially concentrated around the system's ground state

$$\mathcal{C}_0 = \bigcup_{i \in \arg \min_j E_j} \mathcal{C}_i \quad (4.24)$$

that is, the components of crit f with minimal energy. Theorem 4.7 makes this precise.

⁵ Independent of $\varepsilon, \eta, \delta$.

Theorem 4.7. Let μ_∞^η be an invariant measure for (SGD), There exists a constant⁶ $c > 0$ such that

$$\mu_\infty^\eta(\mathcal{C}_0^\delta) \geq 1 - e^{-c/\eta}, \quad \text{where} \quad \mathcal{C}_0^\delta := \bigcup_{i \in \arg \min_j E_j} \mathcal{C}_i^\delta. \quad (4.25)$$

Theorem 4.7 is the latest of our four main results: it asserts that, regardless of how many local minima f possesses, the invariant measure μ_∞^η assigns all but an exponentially small fraction of its mass to a δ -neighborhood of the ground state $\mathcal{C}_0$.

In words, the theorems in this chapter provide the following quantification of the limiting distribution of x_n : in the long run

- (a) the critical region of f is visited exponentially more often than any non-critical region of f (Theorem 4.6);
- (b) in particular, the iterates of (SGD) are exponentially concentrated around the problem's ground state (Theorem 4.7);
- (c) among the mass that remains, every component of f gets a fraction that is exponentially proportional to its energy (Theorem 4.3); and, finally
- (d) every non-minimizing component is “dominated” by a minimizing component that is visited exponentially more often (Theorem 4.4).

Importantly, the problem's energy landscape is shaped by f , but not f alone: the statistics of (SFO) play an equally important role, so we may have $\mathcal{C}_0 \neq \arg \min f$; we discuss this issue in detail in the next section.

4.2 APPLICATIONS AND IMPLICATIONS

4.2.1 SGD with Gaussian noise

To provide some intuition on the energy levels E_i and the resulting Gibbs distribution, we examine the case of SGD with Gaussian noise, a common model for the stochastic gradient noise in deep learning (Mori et al., 2022, Xie et al., 2020).

In the Gaussian case, when the variance depends on x only through the loss value, the energies E_i indeed admit explicit closed-form expressions. For technical reasons,⁷ we work with truncated Gaussian noise, but the same computations apply to the untruncated case and yield the same result.

Assume that $U(x, \omega)$ follows a centered Gaussian distribution with covariance $\sigma^2(f(x))I_d$ conditioned on remaining in $\overline{\mathbb{B}}(0, R(x))$, for some continuous $\sigma^2 : \mathbb{R} \rightarrow (0, +\infty)$ and radius $R(x) > 0$. Akin to V_∞ in Definition 3.8, a key role is played by the function $V : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$V(x) := 2\alpha(f(x)) \quad \text{with} \quad \alpha : \mathbb{R} \rightarrow \mathbb{R} \quad \text{a primitive of} \quad 1/\sigma^2. \quad (4.26)$$

This potential V determines the energy levels E_i . Denote by V_i the value of V on $\mathcal{K}_i$ (recall that f is constant on $\mathcal{K}_i$, so V is as well). The following proposition shows that, up to a constant independent of i , the energy levels are given by the potential values at the critical components.

⁶ Independent of $\varepsilon, \eta, \delta$.

⁷ Formally, standard Gaussian noise is not almost surely bounded and as such would not satisfy Assumption 3.3(b).

Proposition 4.8. *For any $\varepsilon > 0$, if $R(x)$ is large enough depending on $\nabla f(x)$ and ε , then, up to a constant,*

$$E_i = V_i + \mathcal{O}(\varepsilon) \quad \text{for all } i = 1, \dots, K. \quad (4.27)$$

To give some intuition about the proof, we consider the non-truncated Gaussian case for simplicity, but the same argument applies to the truncated case. Since the noise is Gaussian, the Lagrangian and Hamiltonian (from § 3.3.2) have explicit expressions: for every $x, p, v \in \mathbb{R}^d$,

$$\mathcal{H}(x, p) = -\langle \nabla f(x), p \rangle + \log \mathbb{E} \left[e^{\langle p, U(x, \omega) \rangle} \right] \quad (4.28a)$$

$$= -\langle \nabla f(x), p \rangle + \frac{1}{2} \sigma^2(f(x)) \|p\|^2 \quad (4.28b)$$

$$\mathcal{L}(x, v) = \mathcal{H}^*(x, v) = \frac{\|v + \nabla f(x)\|^2}{2\sigma^2(f(x))}. \quad (4.28c)$$

This expression of $\mathcal{L}$ makes it clear that the action functional penalizes the deviation of a path from the flow: for a path $\gamma \in \mathcal{C}([0, T])$ for some $T > 0$, the action cost of γ is given by

$$\mathcal{S}_T(\gamma) = \int_0^T \frac{\|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))} dt. \quad (4.29)$$

The computation of the E_i relies on the following observation. Take a path $\gamma \in \mathcal{C}([0, T])$ for some $T > 0$ and consider φ defined by $\varphi_t = \gamma_{T-t}$ for $t \in [0, T]$. Then, the action cost of φ is given by,

$$\begin{aligned} \mathcal{S}_T(\varphi) &= \int_0^T \frac{\|-\dot{\gamma}_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))} dt \\ &= \int_0^T \frac{\|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))} dt - \int_0^T \frac{2\langle \dot{\gamma}_t, \nabla f(\gamma_t) \rangle}{\sigma^2(f(\gamma_t))} dt \\ &= \mathcal{S}_T(\gamma) - \int_0^T \langle \dot{\gamma}_t, \nabla V(\gamma_t) \rangle dt, \end{aligned} \quad (4.30)$$

since $\nabla V(x) = 2\alpha'(f(x))\nabla f(x)$. Therefore, we get that

$$\mathcal{S}_T(\varphi) = \mathcal{S}_T(\gamma) - (V(\gamma_T) - V(\gamma_0)). \quad (4.31)$$

Take $i, j \in I$. This equality then translates to a relation between $\mathcal{B}_{i,j}$ and $\mathcal{B}_{j,i}$: considering $\gamma \in \mathcal{C}([0, T])$ such that $\gamma_0 \in \mathcal{K}_i$, $\gamma_T \in \mathcal{K}_j$ and taking the infimum over all such paths, we get that

$$\mathcal{B}_{j,i} \leq \mathcal{B}_{i,j} + (V_j - V_i). \quad (4.32)$$

where, since f is constant on $\mathcal{K}_i$ and $\mathcal{K}_j$, we denote by V_i and V_j the values of V on $\mathcal{K}_i$ and $\mathcal{K}_j$ respectively. Reversing the roles of i and j and applying the same argument shows that this inequality is an equality:

$$\mathcal{B}_{j,i} + V_j = \mathcal{B}_{i,j} + V_i. \quad (4.33)$$

Denote by $C_{i,j}$ this common value. Crucially, $C_{i,j}$ is symmetric in i and j .

Now let t^* be a minimum weight spanning tree on the complete but now undirected graph on I with weights $(C_{i,j})_{i,j}$. We show that the minima in the E_i are attained t^* , or more precisely, a directed version of it.

Fix i and consider $t \in \mathcal{T}_i$ to be a spanning tree rooted at i . We have that, since any node j is the origin of exactly one edge in t ,

$$\begin{aligned} \sum_{(j \rightarrow l) \in t} \mathcal{B}_{j,l} + \sum_{j \in I} V_j &= \sum_{(j \rightarrow l) \in t} (\mathcal{B}_{j,l} + V_j) + V_i \\ &= \sum_{(j \rightarrow l) \in t} C_{j,l} + V_i. \end{aligned} \quad (4.34)$$

But, by definition of t^* , this sum is at least equal $\sum_{(j \leftrightarrow l) \in t^*} C_{j,l}$ so we have

$$\sum_{(j \rightarrow l) \in t} \mathcal{B}_{j,l} + \sum_{j \in I} V_j \geq \sum_{(j \leftrightarrow l) \in t^*} C_{j,l} + V_i, \quad (4.35)$$

and by choosing t_i as an oriented version of t^* rooted at i , the equality is attained. Therefore, we have that

$$E_i = \sum_{(j \rightarrow l) \in t_i} \mathcal{B}_{j,l} = \sum_{(j \leftrightarrow l) \in t^*} C_{j,l} - \sum_{j \in I} V_j + V_i, \quad (4.36)$$

or, in short, $E_i = V_i + c$ where $c > 0$ is independent of i .

Therefore, the mass distribution over critical points is governed by a Gibbs measure with potential V , proving [Proposition 4.8](#).

The truncated case is more technical but the same argument applies. It relies on the following lemma, which shows that the Hamiltonian and Lagrangian of the truncated Gaussian are close to those of the standard Gaussian.

Lemma 4.9 ([Lemma A.5.2](#)). *Let $X_R \sim \mathcal{N}(0, \sigma^2 I_d)$ conditioned on being in $\overline{\mathbb{B}}(0, R)$, with cumulant generating function $\tilde{\mathcal{H}}(p) := \log \mathbb{E}[e^{\langle p, X_R \rangle}]$ and Lagrangian $\tilde{\mathcal{L}} := \tilde{\mathcal{H}}^*$. Fix $\varepsilon > 0$ and assume that R/σ is large enough depending on ε . Then, for every $p, v \in \mathbb{R}^d$ such that $\|p\| \leq R/(2\sigma^2)$ and $\|v\| \leq R/4$,*

$$(1 - \varepsilon) \frac{\sigma^2 \|p\|^2}{2} \leq \tilde{\mathcal{H}}(p) \leq (1 + \varepsilon) \frac{\sigma^2 \|p\|^2}{2}, \quad (4.37a)$$

$$(1 - \varepsilon) \frac{\|v\|^2}{2\sigma^2} \leq \tilde{\mathcal{L}}(v) \leq (1 + \varepsilon) \frac{\|v\|^2}{2\sigma^2}. \quad (4.37b)$$

Since the Lagrangian of the truncated Gaussian is $(1 \pm \varepsilon)$ -sandwiched around its untruncated counterpart, the action of any path inherits the same multiplicative approximation. The quasi-potentials, and hence the energies, then satisfy $E_i = V_i + \mathcal{O}(\varepsilon)$ after taking the tree infimum.

Now, two choices of σ^2 are of particular interest. The simplest is the *constant-noise* case $\sigma^2(\cdot) = \sigma^2$: the primitive $\alpha(t) = t/\sigma^2$ gives $V = 2f/\sigma^2$, so the energies are simply proportional to the objective values at the critical components: up to a constant,

$$E_i = \frac{2f_i}{\sigma^2} + \mathcal{O}(\varepsilon) \quad \text{and} \quad \mu_\infty^\eta(\mathcal{U}_i) \propto \exp\left(-\frac{2f_i}{\sigma^2}\right). \quad (4.38)$$

[Figure 14](#) provides a numerical illustration of [Theorems 4.3–4.7](#) on a concrete two-dimensional example in that setting.

Another relevant choice is the *linear-noise* case $\sigma^2(f(x)) = \sigma_1^2(f(x) + \sigma_0^2)$: [Mori et al. \(2022\)](#) showed experimentally that, in deep learning models, the variance of the stochastic gradient noise scales linearly with the loss value. Here the

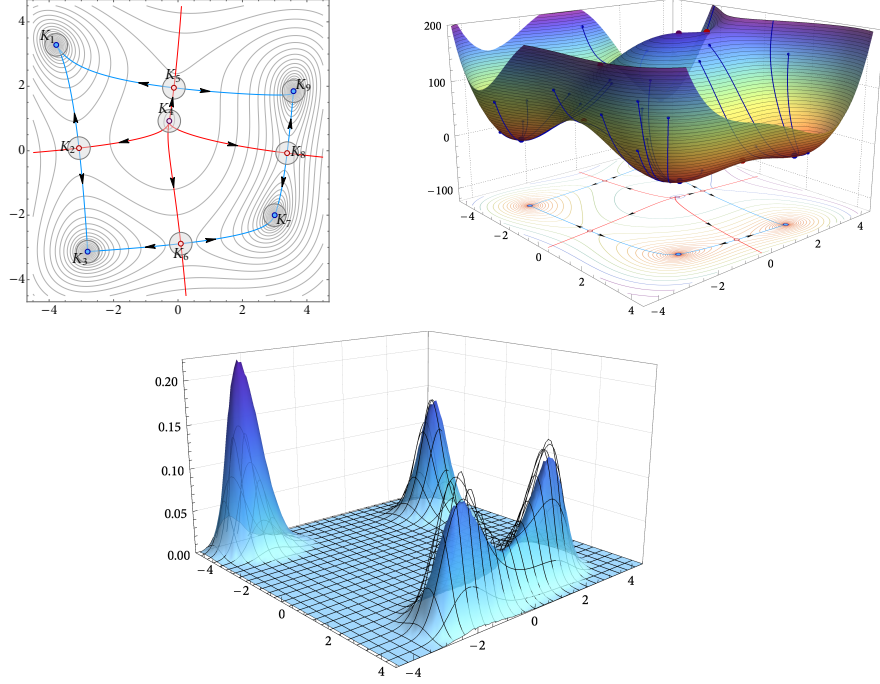


Figure 14: Graphical illustration of [Theorems 4.3–4.7](#) for the Himmelblau test function $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$. The figure to the left depicts the loss landscape of f with several (deterministic) orbits of the gradient flow of f superimposed for visual convenience. The figure in the middle highlights the 9 critical components of f as well as the most likely transitions between them: C_1, C_3, C_7 and C_9 are minimizers (light blue), C_5 is a global maximum (light purple), and the rest are saddle points (light red). The figure to the right illustrates the long-run distribution of 1000 samples of (SGD) run with $\eta = 0.01$ over a horizon of 2×10^4 iterations: the density landscape represents the observed distribution, while the superimposed wireframe indicates our theoretical prediction.

primitive $\alpha(t) = (1/\sigma_1^2) \log(t + \sigma_0^2)$ yields $V_i = (2/\sigma_1^2) \log(f_i + \sigma_0^2)$, a logarithmic dependence on the objective:

$$E_i = \frac{2}{\sigma_1^2} \log(f_i + \sigma_0^2) + \mathcal{O}(\varepsilon) \quad \text{and,} \quad \mu_\infty^\eta(\mathcal{U}_i) \propto (f_i + \sigma_0^2)^{-2/(\sigma_1^2 \eta)}. \quad (4.39)$$

In particular, the linear noise “flattens” the energy landscape, making the selectivity between critical components less pronounced than in the constant-noise case.

Remark 6 (Influence of the batch size). These results allow us to explicit the influence of the batch size B on the energy levels E_i and the resulting Gibbs distribution. For SGD, the scale of the noise is inversely proportional to the batch size: i.e., $\sigma^2(f(x))$ scales as $\frac{1}{B}$. In that case, the values of the potential V_i are multiplied by B so that

$$\frac{E_i}{\eta} \propto \frac{B}{\eta}. \quad (4.40)$$

Interestingly, we recover the linear scaling rules for the learning rate η and batch size B observed in practice: to keep the same behavior when increasing the batch size, the learning rate should be increased proportionally to B ([Goyal et al., 2017](#), [Smith et al., 2018](#)). In our framework, this scaling rule ensures that

the energy levels E_i remain unchanged, so that the long-run distribution of SGD is unaffected by the change in batch size.

In both cases, the ground state coincides with the global minimum of f : [Theorem 4.7](#) then indicates that the iterates of SGD concentrate around the global minimum on the long run.

However, this is not always the case: when σ^2 is allowed to depend on x in a more general way, the energy levels E_i are no longer functions of the objective values f_i , and the ground state may not be the global minimum of f .

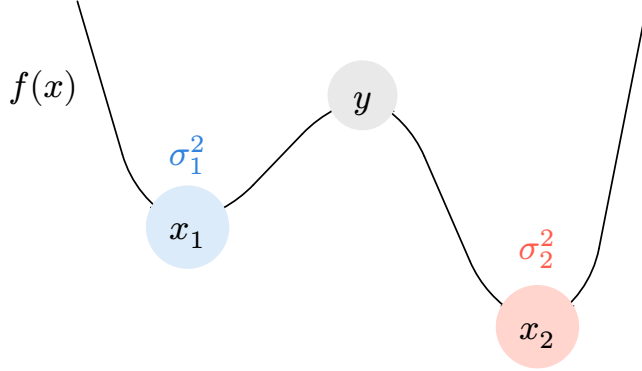


Figure 15: Two-valley landscape with noise variance σ_1^2 in the left basin and σ_2^2 in the right basin. The three critical components are $\mathcal{C}_1 = \{x_1\}$ (left minimum), $\mathcal{C}_2 = \{x_2\}$ (right minimum, with $f(x_2) < f(x_1)$), and $\mathcal{C}_3 = [y - \varepsilon, y + \varepsilon]$ (flat plateau, value $f(y)$).

Consider the function f depicted in [Figure 15](#), with three critical components: $\mathcal{C}_1 = \{x_1\}$ and $\mathcal{C}_2 = \{x_2\}$ are the two local minima, with $f(x_1) > f(x_2)$, and $\mathcal{C}_3 = [y - \varepsilon, y + \varepsilon]$ is a flat plateau of local maxima with constant value $f(y)$. Assume Gaussian noise with variance $\sigma_1^2 > 0$ in the basin of x_1 and $\sigma_2^2 > 0$ in the basin of x_2 , interpolating smoothly in between.

We compute the quasi-potentials $\mathcal{B}_{i,j}$ by applying the symmetry argument of the proof sketch locally in each basin. Since the gradient flow descends from the plateau to each minimum, downhill transitions have zero action:

$$\mathcal{B}_{3,1} = \mathcal{B}_{3,2} = 0. \quad (4.41)$$

Let $\Delta_1 := f(y) - f(x_1)$ and $\Delta_2 := f(y) - f(x_2)$ denote the *depths* of x_1 and x_2 , measuring how far each local minimum lies below the plateau. Uphill, the path from x_1 to the plateau stays in the left basin (noise σ_1^2) and the path from x_2 stays in the right basin (noise σ_2^2), giving

$$\mathcal{B}_{1,3} = \frac{2\Delta_1}{\sigma_1^2}, \quad \mathcal{B}_{2,3} = \frac{2\Delta_2}{\sigma_2^2}. \quad (4.42)$$

Crossing between the two minima costs at most a passage through the plateau, so $\mathcal{B}_{1,2} = \mathcal{B}_{1,3}$ and $\mathcal{B}_{2,1} = \mathcal{B}_{2,3}$. The full matrix is thus

$$(\mathcal{B}_{i,j})_{i,j} = \begin{pmatrix} 0 & \frac{2\Delta_1}{\sigma_1^2} & \frac{2\Delta_1}{\sigma_1^2} \\ \frac{2\Delta_2}{\sigma_2^2} & 0 & \frac{2\Delta_2}{\sigma_2^2} \\ 0 & 0 & 0 \end{pmatrix}. \quad (4.43)$$

Since exit costs from $\mathcal{C}_3$ are zero (third row), spanning trees rooted at $\mathcal{C}_1$ or $\mathcal{C}_2$ route through $\mathcal{C}_3$ at no cost and only require the edge from the other local minimum. Computing the minimum spanning trees, the energy levels are

$$E_1 = \frac{2\Delta_2}{\sigma_2^2}, \quad E_2 = \frac{2\Delta_1}{\sigma_1^2}, \quad E_3 = E_1 + E_2. \quad (4.44)$$

Note that $E_3 = E_1 + E_2 > \max(E_1, E_2)$, so $\mathcal{C}_3$ is never the ground state — reflecting the instability of local maxima.

Constant noise ($\sigma_1 = \sigma_2 = \sigma$). In this case $E_1 = 2\Delta_2/\sigma^2$ and $E_2 = 2\Delta_1/\sigma^2$. Since $f(x_2) < f(x_1)$, we have $\Delta_2 > \Delta_1$ and hence $E_1 > E_2$, so the ground state is $\mathcal{C}_2$, the global minimum of f . This recovers (4.38) and confirms that, in the constant-noise case, the ground state is always the deepest minimum.

Non-constant noise ($\sigma_1 \neq \sigma_2$). When the noise varies, (4.44) shows that $\mathcal{C}_1$ is the ground state if and only if

$$\frac{\Delta_2}{\sigma_2^2} < \frac{\Delta_1}{\sigma_1^2}, \quad (4.45)$$

which can hold even when $f(x_2) < f(x_1)$, provided σ_1 is small enough relative to σ_2 .

Mori et al. (2022) and Xie et al. (2020) showed experimentally that, in deep learning models, the noise covariance near a minimum is approximately proportional to the Hessian of the loss; in our one-dimensional example, this means $\sigma_1^2 \propto f''(x_1)$ and $\sigma_2^2 \propto f''(x_2)$. Hence, a flat minimum x_1 — one with small curvature $f''(x_1)$ — has small noise variance σ_1^2 , which raises the exit cost $\mathcal{B}_{1,3} = 2\Delta_1/\sigma_1^2$ and hence E_2 , making $\mathcal{K}_1$ the ground state even if $f(x_1) > f(x_2)$. This is a quantitative manifestation of the flat-minima bias of SGD.

The next section gives a general analysis of this phenomenon.

4.2.2 Local dependencies

The modeling of the noise is crucial to the understanding of the dynamics of (SGD). This has been often underlined, especially in the exit-time literature (Hu et al., 2019, Xie et al., 2020, Mori et al., 2022, Jastrzębski et al., 2017)

In particular, Mori et al. (2022), after heuristic arguments and empirical investigation, suggested the following model for the noise in deep learning training close to a minimum $\mathcal{C}_i = \{x^*\}$: for x close to $\mathcal{C}_i$,

$$U(x) = \mathcal{N}(0, \sigma^2 f(x) H^*), \quad (4.46)$$

where $H^* := \text{Hess } f(x^*)$ is the Hessian of f at x^* and $\sigma^2 > 0$ is a constant.

We explain here how this model of the noise influences the invariant measure of (SGD). We assume that a weaker version of (4.46) holds, namely that the cumulant-generating function (CGF) of the noise satisfies

$$\log \mathbb{E} e^{\langle v, U(x; \omega) \rangle} \leq \frac{\sigma^2 f(x)}{2} \langle v, H^* v \rangle \quad \text{for all } x \text{ near } \mathcal{C}_i, \quad (4.47)$$

with f remaining positive near $\mathcal{C}_i$.

We also follow the structure assumption of [Mori et al. \(2022\)](#): assume that $\log f$ is separable in the eigenbasis of H^* :

$$\log f(x) = \sum_{\lambda \in \text{eig } H^*} g^\lambda(x_\lambda) \quad (4.48)$$

where x_λ denotes the projection of x on the eigenspace of λ .

Lemma 4.10. *Assuming (4.47) and (4.48), for all $j \neq i$, we have*

$$E_j \geq 2 \min \left\{ \sum_{\lambda \in \text{eig } H^*} \frac{g^\lambda(x_\lambda) - g_i^\lambda}{\lambda \sigma^2} : x, \text{dist}(x, C_i) = \delta \right\} \quad (4.49)$$

where g_i^λ is the value of g^λ on K_i .

This result shows that the energy of each component of $\text{crit}(f)$ is lower bounded by the right-hand side (RHS) of (4.49). This quantity scales as the reciprocal of the eigenvalues of H^* so, as the minimum becomes flatter (i.e., the eigenvalues of H^* become smaller), the energy levels of all other components become larger: thus, relative to component C_i , the other components all become less probable. Moreover, note that the RHS of (4.49) only scales logarithmically with the value of the objective function around C_i , i.e., the depth of the minimum: this means that the “flatness” of the minimum plays a greater role in the relative probabilities of the components than the depth.

This also shows that deepest minima do not necessarily correspond to the ground state of the problem: if σ^2 or the eigenvalues of H^* is small enough compared to the noise level outside the δ -neighborhood of C_i , then C_i will be the ground state of the system even if it is not the deepest minimum; we provide a formal proof of this in [Appendix A.4.3](#).

4.2.3 Convergence of empirical measures

Theorems [Theorem 4.3–Theorem 4.7](#) hold for the invariant measure μ_∞^η , which requires SGD to have reached stationarity. In practice, SGD is run for a finite number of iterations N ; the natural quantity is then the *mean occupation measure*

$$\mu_N(\mathcal{B}) := \mathbb{E} \left[\frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}\{x_n \in \mathcal{B}\} \right] \quad (4.50)$$

for every Borel set $\mathcal{B} \subseteq \mathbb{R}^d$. In words, $\mu_N(\mathcal{B})$ measures the mean fraction of time that SGD has spent in $\mathcal{B}$ up to time N . To extend our results to μ_N , we require a mild strengthening of [Assumption 3.4](#):

Assumption 4.5 (Variant of [Assumption 3.4](#)). The signal-to-noise ratio of G satisfies

$$\frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \rightarrow \infty \quad \text{as} \quad \|x\| \rightarrow \infty. \quad (\text{SNR}^{**})$$

Under [Assumption 4.5](#), the sequence $(\mu_N)_{N \geq 0}$ is tight⁸; any weak limit point is an invariant measure, so [Theorem 4.3](#), [Theorem 4.4](#), [Theorem 4.6](#), and

⁸ Here, tightness means that for every $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset \mathbb{R}^d$ such that $\mu_N \geq 1 - \varepsilon$ for all sufficiently large N . In other words, the occupation measures do not lose mass at infinity. See e.g., [Kallenberg \(2021, Chap. 23\)](#) for details.

[Theorem 4.7](#) apply to any such limit point. Since those results hold uniformly over sufficiently small neighborhoods and step sizes, passing to the limit in N yields the following occupation-measure analogues.

Theorem 4.3* (Occupation variant of [Theorem 4.3](#)). *For N large enough, we have*

$$\exp\left(-\frac{E_i - \min_{j \in I} E_j}{\eta} - \frac{\varepsilon}{\eta}\right) \leq \mu_N(\mathcal{U}_i) \leq \exp\left(-\frac{E_i - \min_{j \in I} E_j}{\eta} + \frac{\varepsilon}{\eta}\right). \quad (4.51)$$

More compactly, with notation as above, we have:

$$\mu_N(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right). \quad (4.52)$$

Theorem 4.4* (Occupation variant of [Theorem 4.4](#)). *Let $\mathcal{C}_i$ be a non-minimizing component of f . Then, there exists a minimizing component $\mathcal{C}_j$ of f and a positive constant⁹ $c > 0$ such that, for N large enough,*

$$\frac{\mu_N(\mathcal{C}_i^\delta)}{\mu_N(\mathcal{C}_j^\delta)} \leq \exp\left(-\frac{c + \varepsilon}{\eta}\right). \quad (4.53)$$

Theorem 4.6* (Occupation variant of [Theorem 4.6](#)). *There exists a constant¹⁰ $c > 0$ such that, for N sufficiently large,*

$$\mu_N(\mathcal{C}_{\min}^\delta) \geq 1 - e^{-c/\eta}, \quad \text{where } \mathcal{C}_{\min}^\delta := \bigcup_{i: \mathcal{K}_i \text{ minimizing}} \mathcal{C}_i^\delta \quad (4.54)$$

and, therefore,

$$\mu_N(\mathcal{C}^\delta) \geq 1 - e^{-c/\eta}, \quad \text{where } \mathcal{C}^\delta = \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta. \quad (4.55)$$

Theorem 4.7* (Occupation variant of [Theorem 4.7](#)). *There exists a constant¹¹ $c > 0$ such that, for N sufficiently large,*

$$\mu_N(\mathcal{U}_0) \geq 1 - e^{-c/\eta}. \quad (4.56)$$

4.2.4 Dissipative systems

The results of the preceding section hold for any initial condition and require only mild assumptions on the noise and the objective. In many applications, however, the objective satisfies a stronger structural condition known as *dissipativity*, which is standard in the SDE and sampling literature ([Erdogdu et al., 2018](#), [Li and Wang, 2022](#), [Raginsky et al., 2017](#), [Lu et al., 2021](#)). We examine this setting in detail, following [Example 3.2](#). Dissipativity has two types of consequences. First, it implies tightness of the occupation measure sequence, so [Theorem 4.3*](#), [Theorem 4.4*](#), [Theorem 4.6*](#), and [Theorem 4.7*](#) continue to hold without any additional assumption. Second, and more importantly, dissipativity enables integration of polynomially growing functions against the occupation measure — not just bounded test functions. This yields three concrete corollaries that connect our theoretical results to observed phenomena in deep learning.

⁹ Independent of $\varepsilon, \eta, \delta$.

¹⁰ Independent of $\varepsilon, \eta, \delta$.

¹¹ Independent of $\varepsilon, \eta, \delta$.

The formal definition of dissipativity is as follows.

Assumption 4.6 (Dissipativity). There exist constants $\alpha, \beta > 0$ such that, for all $x \in \mathbb{R}^d$,

$$\langle \nabla f(x), x \rangle \geq \alpha \|x\|^2 - \beta. \quad (4.57)$$

[Assumption 4.6](#) is a coercivity condition on the gradient drift: as $\|x\| \rightarrow \infty$, the gradient $\nabla f(x)$ points increasingly inward, preventing the iterates from escaping to infinity. In particular, it implies $f(x) \rightarrow +\infty$ as $\|x\| \rightarrow \infty$ and that $\|\nabla f(x)\| = \Omega(\|x\|)$ outside a compact set.

Beyond tightness, dissipativity ensures that the iterates do not stray far from the origin: the occupation measure μ_n satisfies uniform polynomial moment bounds, so for any continuous function ϕ growing at most polynomially, $\int \phi d\mu_n$ is well defined and converges as $N \rightarrow +\infty$. The three results below exploit this integrability.

Convergence of the loss value

[Zhang et al. \(2022b\)](#) observed experimentally that, in large-scale neural network training, the loss stabilizes even though the weights do not converge to stationary points. The following proposition provides a theoretical grounding for this phenomenon: under dissipativity and a mild condition on the loss values at the ground state, the time-averaged loss converges to the global minimum up to an exponentially small correction.

Proposition 4.15. Under [Assumption 4.6](#), assume that

$$\left\{ f(x) : x \in \mathcal{K}_i, i \in \arg \min_j E_j \right\} = \{f^*\} \quad (4.58)$$

i.e., it is a singleton. For any $\varepsilon > 0$, for neighborhoods $\mathcal{U}_i$ small enough and η small enough, for N large enough:

$$\mathbb{E} \left[\frac{1}{N} \sum_{n=0}^{N-1} f(x_n) \right] = f^* + \mathcal{O}(e^{-c/\eta}) \quad (4.59)$$

for some $c > 0$.

Indeed, by [Theorem 4.7*](#), μ_N concentrates on the ground state $\mathcal{C}_0$ with complement probability at most $e^{-c/\eta}$. To conclude, it suffices to use that $f = f^*$ on $\mathcal{C}_0$ and dissipativity to control the contribution from the complement of $\mathcal{C}_0$. In other words, the running average of the loss converges to the globally minimal loss value, with corrections of order $e^{-c/\eta}$ that are negligible for small step-size η .

Weight averaging

Classical convex optimization theory typically emphasizes the theoretical benefits of the running average of the iterates ([Ruppert, 1988](#), [Polyak and Juditsky, 1992](#)) compared to the last iterate. Though a center piece of convergence analysis for convex optimization – and more broadly monotone variational inequalities – the running average is often dismissed in the non-convex setting: averaging the iterates does not make any sense in theory in the non-convex setting. However, there has been a renewed interest in weight averaging

techniques in deep learning, which have been empirically shown to improve generalization and robustness (Izmailov et al., 2018, Athiwaratkun et al., 2019, Nikishin et al., 2018, Yang et al., 2019, Garipov et al., 2018).

Thus, the running average

$$\bar{x}_N = \frac{1}{N} \sum_{n=0}^{N-1} x_n \quad (4.60)$$

of the iterates is used as the final model rather than the last iterate. This average tends to converge to flatter regions of the loss landscape. The following proposition provides a theoretical justification in the non-convex setting: when the components minimizing the energy are convex, this average actually converges to the ground state.

Proposition 4.16. *Under Assumption 4.6, assume that $\arg \min_j E_j$ is a singleton $\{i\}$ and that $\mathcal{K}_i$ is convex. For any $\varepsilon > 0$, for neighborhoods $\mathcal{U}_i$ small enough and η small enough, for N large enough:*

$$\mathbb{E}[\text{dist}(\bar{x}_N, \mathcal{C}_0)] \leq \varepsilon + \mathcal{O}(e^{-c/\eta}), \quad (4.61)$$

where $\bar{x}_N := \frac{1}{N} \sum_{n=0}^{N-1} x_n$ and $c > 0$.

By convexity of each $\mathcal{K}_i$, Jensen's inequality gives

$$\mathbb{E}[\text{dist}(\bar{x}_N, \mathcal{K}_i)] \leq \mathbb{E}\left[\frac{1}{N} \sum_{n=0}^{N-1} \text{dist}(x_n, \mathcal{K}_i)\right] = \int \text{dist}(\cdot, \mathcal{K}_i) d\mu_N \quad (4.62)$$

Since $d(\cdot, \mathcal{K}_i)$ grows linearly, dissipativity makes this integral well defined. Theorem 4.7* then implies that μ_N assigns mass at least $1 - e^{-c/\eta}$ to $\mathcal{U}_0$, the ε -neighborhood of $\mathcal{C}_0$; hence the integral is at most $\varepsilon + \mathcal{O}(e^{-c/\eta})$. This results thus shows that, in expectation, the running average of the iterates is actually very close to the ground state $\mathcal{C}_0$. While the iterates of SGD themselves wander around different components and are close to the ground states only with high probability, the running average on the contrary actually converges to the ground state. This shows how averaging in the non-convex setting can be seen as a principled technique to obtain a point of the ground states. This thus provides a theoretical justification for the empirical success of weight averaging techniques in deep learning, even in the non-convex setting.

Online convex optimization as a proxy for deep learning dynamics

A striking empirical observation in deep learning is that online convex optimization (OCO) theory very accurately describes the behavior of gradient-based optimizers, despite the non-convexity of the loss landscape (Schaipp et al., 2025, Defazio et al., 2023, 2024, Ahn et al., 2024). We now sketch one possible theoretical justification via dissipativity.

Define the ground state set $\mathcal{C}_0 := \bigcup_{i \in \arg \min_j E_j} \mathcal{K}_i$ and, for each iterate x_n , its projection $p_n := \text{proj}_{\mathcal{C}_0}(x_n)$ onto $\mathcal{C}_0$. The excess loss admits the following elementary decomposition:

$$\frac{1}{N} \sum_{n=0}^{N-1} [f(x_n) - f(p_n)] = \frac{1}{N} \sum_{n=0}^{N-1} \langle \hat{g}_n, x_n - p_n \rangle \quad (4.63a)$$

$$+ \frac{1}{N} \sum_{n=0}^{N-1} [f(x_n) - f(p_n) - \langle \hat{g}_n, x_n - p_n \rangle], \quad (4.63b)$$

where $\hat{g}_n = \nabla f(x_n) + U(x_n, \omega_n)$ is the stochastic gradient at step n . The first term on the right-hand side is the standard OCO regret term; standard online gradient descent analysis bounds it as

$$\frac{1}{N} \sum_{n=0}^{N-1} \langle \hat{g}_n, x_n - p_n \rangle \leq \frac{\|x_0 - p_0\|^2}{2\eta N} + \frac{\eta}{2N} \sum_{n=0}^{N-1} \|\hat{g}_n\|^2, \quad (4.64)$$

which vanishes as $N \rightarrow \infty$ for fixed η . The second term in (4.63b) captures the *average non-convexity* of f along the path. [Tran et al. \(2024\)](#) empirically study this quantity and observe that this term is negative along typical SGD trajectories in deep learning; the following proposition provides a theoretical justification for this observation under dissipativity and a local convexity condition near $\mathcal{C}_0$.

Proposition 4.17. *Under [Assumption 4.6](#), assume that f is “convex near $\mathcal{C}_0$ ” in the following sense: in a neighborhood of $\mathcal{C}_0$,*

$$f(x) + \langle \nabla f(x), \text{proj}_{\mathcal{C}_0}(x) - x \rangle \leq f(\text{proj}_{\mathcal{C}_0}(x)). \quad (4.65)$$

For any $\varepsilon > 0$, for neighborhoods $\mathcal{U}_i$ small enough and η small enough, for N large enough:

$$\frac{1}{N} \sum_{n=0}^{N-1} [f(x_n) - f(p_n) - \langle g_n, x_n - p_n \rangle] \leq \mathcal{O}(e^{-c/\eta}) \quad (4.66)$$

for some $c > 0$.

The local convexity assumption states that f satisfies the sub-gradient inequality with respect to the projection $\text{proj}_{\mathcal{C}_0}$ near $\mathcal{C}_0$: the non-convexity

$$f(x) - f(p) - \langle \nabla f(x), p - x \rangle \quad (4.67)$$

is non-positive close to $\mathcal{C}_0$. Since the same quantity grows at most polynomially by dissipativity, the contribution from iterates far from $\mathcal{C}_0$ is uniformly integrable against μ_N . By [Theorem 4.7*](#), SGD spends all but a $e^{-c/\eta}$ fraction of its time near $\mathcal{C}_0$, where the non-convexity is non-positive; hence the average is $\mathcal{O}(e^{-c/\eta})$.

Taken together, [Proposition 4.17](#) and the standard OCO bound imply that, once η is small and N is large, the excess loss (4.63b) is driven solely by the standard OCO regret term. In other words, the non-convex SGD dynamics are well-approximated by a convex OCO algorithm operating near the ground state, providing a theoretical basis for the empirically observed agreement between OCO theory and gradient-based optimizers in deep learning.

Towards non-asymptotic occupation-measure bounds

The above results are asymptotic in n : they guarantee that μ_n eventually concentrates on the ground state, but they do not quantify the rate of convergence. One natural approach would be to study the mixing time of the induced chain, whose transition probabilities satisfy $p_{ij} \propto e^{-\mathcal{B}_{ij}/\eta}$. However, using the mixing time of the induced chain can be overly pessimistic. Consider the two-valley example of [Figure 15](#) with $f(x_2) \ll f(x_1), f(y)$ and $\sigma_1 = \sigma_2$. The induced chain mixes very slowly because $f(y) - f(x_1)$ is large; yet the probability of being

in a neighborhood of x_2 is very high even for moderate n , because the gap $f(x_1) - f(x_2)$ is large. A prerequisite for $\mu_n(\mathcal{U}_0)$ to reach its asymptotic value is for SGD to visit $\mathcal{U}_0$ at least once. We will therefore sharpen these results in the next chapter by studying the first hitting time of $\mathcal{U}_0$ directly, which yields the non-asymptotic occupation-measure bounds of [§ 4.2.3](#).

THE GLOBAL CONVERGENCE TIME OF SGD

5.1	An instructive example	71
5.2	Hitting-time analysis	75
5.2.1	Hitting time for the induced chain	75
5.2.2	Refined setup	77
5.2.3	Hitting time for accelerated SGD	80
5.2.4	Global hitting time of SGD	82
5.2.5	Basin-dependent global hitting time of SGD	83
5.3	Extensions and refinements	85
5.3.1	No spurious local minima	85
5.3.2	Influence of the loss landscape	85
5.3.3	Numerical illustrations	88
5.3.4	Non-asymptotic occupation measures	90

This chapter presents the main results on the global convergence time of (SGD); the material is based on [Azizian et al. \(2025a\)](#) but completely overhauled, highlighting in particular how the results follow directly from the essential ingredients of [Chapter 4](#). In [§ 5.1](#), we introduce the key ideas through a two-dimensional worked example, giving a concrete preview of the main results before the full analysis. In [§ 5.2](#), we carry out the full rigorous analysis: we first derive a combinatorial formula for the expected hitting time of a finite Markov chain and adapt the quasi-potential framework to first-passage rather than recurrent behavior; we then combine these ingredients to obtain a characterization of the global convergence time of SGD, culminating in [Theorems 5.7](#) and [5.10](#). In [§ 5.3](#), we provide a series of refinements and extensions, in particular linking the geometry of the loss landscape to the global convergence time.

5.1 AN INSTRUCTIVE EXAMPLE

Stating our results in their most general form requires some fairly involved technical apparatus, so we begin with a warm-up example intended to introduce basic concepts and develop intuition for the sequel. Specifically, we give a high-level overview of the main results for a simple two-dimensional example which is easy to plot and visualize. The material here is presented at an informal level; the rigorous treatment is deferred to [§ 5.2](#) below.

The example we work with is a modified version of the well-known “three-hump camel” test function, as detailed in [Figure 16](#). This is a multimodal function with five critical points, indexed p_1 through p_5 : two are saddles (p_2 and p_4), three are minimizers (p_1 , p_3 , and p_5), and the global minimum is attained at $p_1 \approx (-2.573, 1.029)$. We further assume that (SGD) is run with stochastic gradients of the form $\hat{g}_n = \nabla f(x_n) + Z_n$, where Z_n is an

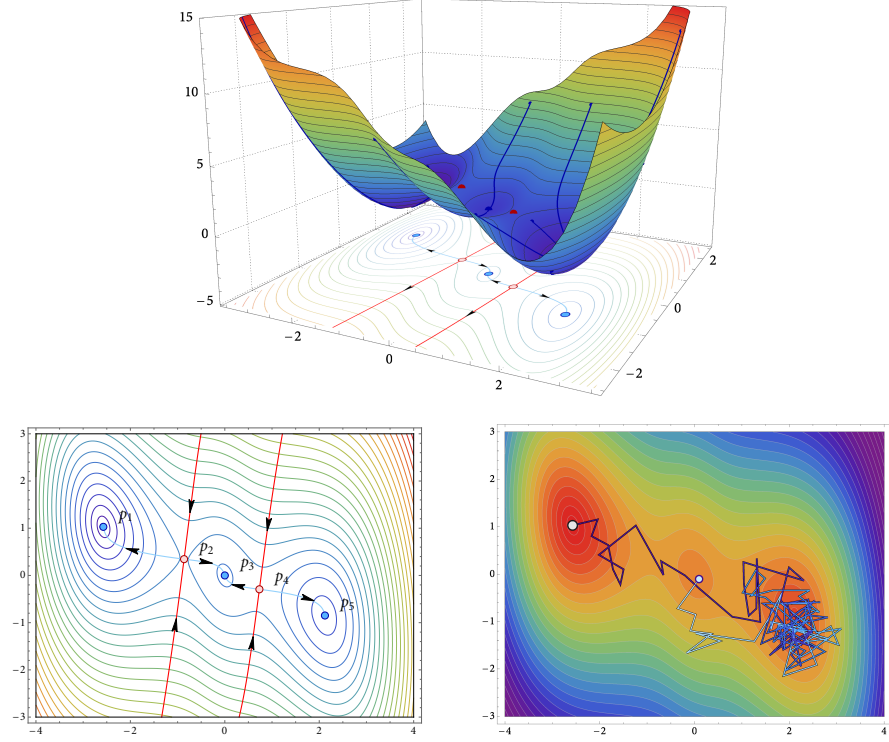


Figure 16: Visualization of the three-hump camel benchmark $f(x_1, x_2) = 2x_1^6/13 + x_1^5/8 - 91x_1^4/64 - 24x_1^3/48 + 42x_1^2/16 + 5x_2^2/4 + x_1x_2$. The figure to the left depicts the loss landscape of f with several (deterministic) gradient flow orbits superimposed for visual convenience. The figure in the middle highlights the 5 critical points of f (light blue for local minimizers, light red for saddle points): the red curves depict the stable manifolds of the saddle points p_2 and p_4 , and they partition the space into three basins of attraction, one per minimizer; the blue curves are gradient flow orbits that realize the edges of the transition graph of f . Finally, the figure to the right illustrates a trajectory of (SGD) starting near p_3 , and stopped when it gets within 10^{-2} of p_1 , the global minimum of f ; color represents time, with darker hues indicating later iterations of (SGD). The plotted trajectory gets to p_1 after first being trapped by p_5 ; the sojourn time from p_5 to p_1 is the leading contribution to the global convergence time of (SGD).

independent and identically distributed (i.i.d.) sequence of Gaussian random vectors with covariance $\sigma^2 I_d$. Then, fixing an initialization $x_0 \leftarrow x \in \mathbb{R}^2$, the global convergence time is

$$\tau = \inf\{n = 0, 1, \dots : \|x_n - p_1\| \leq \delta\} \quad (5.1)$$

for some fixed error margin $\delta > 0$ (e.g., $\delta = 10^{-2}$ in Figure 16).

The core of our analysis may then be summarized as follows.

1. With exponentially high probability, (SGD) spends most of its time near the critical points of f (and, in particular, its minimizers).
2. To leading order, the time that (SGD) takes to reach a neighborhood of $\arg \min f$ is determined by the chain of critical points visited by x_n , and by the average time required to transition from one to the next.

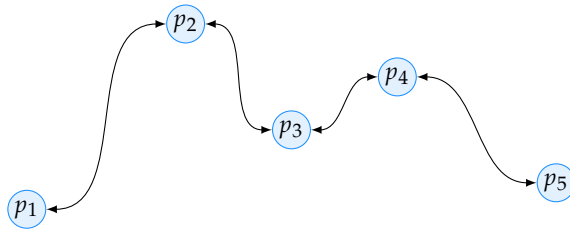
The technical scaffolding required to make this intuition precise is provided by a weighted directed graph which quantifies the difficulty of (SGD) making a direct transition between two critical points of f . In the context of our example, this graph simplifies and is constructed as follows.

- The nodes of the graph are the critical points $p_i, i = 1, \dots, 5$, of f .
- Two nodes are joined by an edge if there exists a solution orbit of the gradient flow $\dot{x}(t) = -\nabla f(x(t))$ whose closure connects said points.¹ Importantly, if $p_i \rightsquigarrow p_j$, we also have $p_j \rightsquigarrow p_i$ by default.
- The weight of an edge $p_i \rightsquigarrow p_j$ is given by

$$B_{ij} = 2[f(p_j) - f(p_i)]_+ / \sigma^2. \quad (5.2)$$

An instructive way of interpreting this expression is as follows: if $f(p_j) \leq f(p_i)$, the transition of (SGD) from p_i to p_j is “costless”; otherwise, if $f(p_j) > f(p_i)$, the noise in (SGD) can still lead to an *ascent* from p_i to p_j , but the cost of such a transition is proportional to the potential difference $f(p_j) - f(p_i)$, and inversely proportional to the variance of the noise.

In our example, this construction yields the path graph below (where, for visual clarity, the height of each node corresponds to its objective value):



The topology of this graph captures the fact that the entire space is partitioned into the basins of attraction of p_1, p_3 and p_5 , as shown in Figure 16. Since p_1 and p_5 are separated by p_3 , there can be no gradient flow orbits joining “non-successive” critical points – e.g., p_1 to p_3 – leading to the path graph structure depicted above.

We fix an initial condition x for (SGD) in the basin of attraction of p_3 . In this case, the most likely event is that (SGD) will first be attracted to p_3 on its way to the global minimum p_1 , so we estimate the time it takes to reach p_1 from

¹ Formally, two distinct critical points p_i, p_j are joined by an edge if there exists a solution orbit $x(t)$ of the gradient flow of f such that $\{p_i, p_j\} = \{\lim_{t \rightarrow \pm\infty} x(t)\}$.

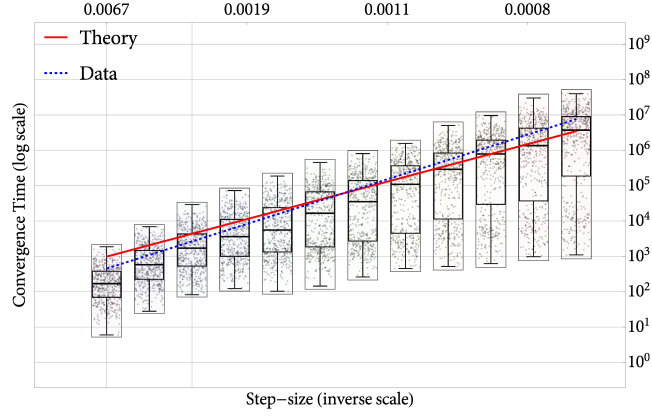


Figure 17: Statistical analysis of the number of iterations τ required for (SGD) to reach the global minimum of the three-hump camel function of Figure 16. For each value of η , we performed 500 runs of (SGD) initialized near p_3 with Gaussian noise ($\sigma = 50$), and we recorded the number of iterations required to reach the vicinity of p_1 (within $\delta = 10^{-2}$). The resulting boxplots are displayed in log-inverse scale, with a distribution chart overlay to illustrate the full empirical distribution of τ for the given value of η . As per (5.3), we see that $E[\tau]$ scales exponentially with $1/\eta$. To measure the goodness-of-fit, the solid “Data” line represents a linear regression fit for our data, while the dashed “Theory” line is a linear fit with slope given by (5.3). The two fits have R^2 values of 0.99 and 0.97 respectively, indicating a strong agreement between theory and experiment.

p_3 . However, this time cannot be determined only by the local geometry of f along the “direct” transition path $p_3 \rightsquigarrow p_2 \rightsquigarrow p_1$: with positive probability, (SGD) may first jump over p_4 and be trapped in p_5 . In that case, the process will first have to escape from p_5 , and then follow the “indirect” transition path $p_5 \rightsquigarrow p_4 \rightsquigarrow p_3 \rightsquigarrow p_2 \rightsquigarrow p_1$. As a result, the mean global convergence time of (SGD) is not affected solely by the obstacles that lie on the most “direct” path from the initialization to the global minimum of f , but also by the traps that the process can fall into along every other “indirect” path as well.

With all this in mind, when instantiated to the example at hand, the analysis of the next section yields the following expression for the mean convergence time of (SGD):

$$E_x[\tau] \approx \exp\left(\frac{2(f(p_2) - f(p_5))}{\eta\sigma^2}\right). \quad (5.3)$$

This expression shows that the global convergence time scales exponentially with the inverse of the step-size and the variance of the noise, and it is dominated by the objective value gap that x_n must clear if it is trapped by p_5 . We verify experimentally the accuracy of this derivation in Figure 17.

To provide some more context, the precise expression for the coefficient of $1/\eta$ in (5.3) – which we denote as $E(p_1|p_3)$ – is obtained by aggregating the graph weights B_{ij} according to the formula

$$\begin{aligned} E(p_1|p_3) &= B_{32} + B_{54} - \min\{B_{34}, B_{54}\} \\ &= \frac{2 \max\{f(p_2) - f(p_3), f(p_2) - f(p_5)\}}{\sigma^2}. \end{aligned} \quad (5.4)$$

This formula comes from (5.8) applied to this specific case. In words, the exponent $E(p_1|p_3)$ scales with the largest value gap that x_n may need to clear

in order to reach $\arg \min f$. In the specific example of Figure 16, we have $f(p_3) > f(p_5)$, so $f(p_2) - f(p_5) > f(p_2) - f(p_3)$, and hence

$$E(p_1|p_3) = 2[f(p_2) - f(p_5)]/\sigma^2. \quad (5.5)$$

This is precisely the obstacle that (SGD) must clear if it gets trapped by p_5 , hence our intuition (cf. Figure 16).

All this may sound paradoxical at first: (5.3) shows that the global convergence time of (SGD) is not driven by the “shortest”, *least costly* chain of transitions leading to $\arg \min f$, but by the “longest”, *most costly* one. In this sense, $E(p_1|p_3)$ directly encodes all non-local information of f that ends up affecting the global convergence time of (SGD), so it can be seen as an explicit measure of the “hardness” of the non-convex landscape: in particular, a slight increase in this gap would have an *exponential impact* on the global convergence time of (SGD), indicating the difficulty of learning even in this simple example.

5.2 HITTING-TIME ANALYSIS

This section presents our main results on the global convergence time of (SGD), culminating in Theorems 5.7 and 5.10. In § 5.2.1, following § 4.1.1, we begin with a general Markov chain on a finite state space, derive a combinatorial formula for the expected hitting time (5.8), and then apply it to the accelerated induced chain. In § 5.2.2, we refine the setup of Chapter 4 for the hitting-time problem — a technical but necessary step, as hitting times require tracking first passages rather than recurrent visits. In § 5.2.3, these ingredients combine to yield matching upper and lower bounds on the expected hitting time of the accelerated induced chain: Proposition 5.1 provides exponential estimates on the transition probabilities; (5.8) converts these into the tree formula for the expected hitting time, expressed in terms of the energy $E(Q|i)$ (Definition 5.1); and Proposition 5.2 controls the duration of each individual transition, ensuring it remains sub-exponential. In § 5.2.4, these bounds are transferred to the original SGD sequence: the upper bound follows directly from Theorem 5.4; the lower bound requires additional work, exploiting the local stability of SGD near each global minimizer. Combining both bounds yields Theorem 5.7, our main characterization: the global convergence time scales as $\exp(E(Q|x_0)/\eta)$, where $E(Q|x_0)$ encodes all non-local information about the landscape. In particular, this exponent is not driven by the shortest, least costly path to $\arg \min f$, but by the longest, most costly one: a deep spurious trap, even if rarely visited, can dominate the global convergence time. Finally, § 5.2.5 sharpens these bounds when the initialization lies in the basin of attraction of a minimizing component $\mathcal{C}_i$: in that case, the exponent simplifies and depends only on the component, not on the precise initialization.

5.2.1 Hitting time for the induced chain

As in § 4.1.1, we begin with a general Markov chain on a finite state space, derive a combinatorial formula for the quantity of interest, and then apply it to the accelerated induced chain.

We are for now ignoring the technical differences with Chapter 4 which are required make the derivations below rigorous, and will be handled in § 5.2.2.

Consider a Markov chain on a state space $\mathcal{V} = \{1, \dots, N_{\text{crit}}\}$, with transition probabilities $(p_{ij})_{i,j}$. Given a target set $\mathcal{Q} \subseteq \mathcal{V}$, the *expected hitting time* of $\mathcal{Q}$ from $i \notin \mathcal{Q}$ is

$$h_i := \mathbb{E}_i[\tau_{\mathcal{Q}}] := \mathbb{E}_i[\inf\{n \geq 0 : x_n \in \mathcal{Q}\}], \quad (5.6)$$

with the convention $h_i = 0$ for $i \in \mathcal{Q}$. The vector $(h_i)_{i \in \mathcal{V}}$ is the unique solution to the system:

$$h_i = 0 \quad (i \in \mathcal{Q}), \quad h_i = 1 + \sum_{j \in \mathcal{V}} p_{ij} h_j \quad (i \notin \mathcal{Q}). \quad (5.7)$$

As in the invariant measure case, this linear system could in principle be solved by inverting a matrix, but the resulting formula is not explicit. For our purposes, we need an explicit form.

We first introduce the graph notation. Let $r := \text{card}(\mathcal{V} \setminus \mathcal{Q})$. A $\mathcal{Q}$ -graph on $\mathcal{V}$ is a directed graph with edges from $\mathcal{V} \setminus \mathcal{Q}$ to $\mathcal{V}$ such that: (i) every non-target vertex has exactly one outgoing edge, and (ii) there are no directed cycles. The set of $\mathcal{Q}$ -graphs is denoted $\mathcal{G}(\mathcal{Q})$. Note that in every $g \in \mathcal{G}(\mathcal{Q})$, every non-target vertex i has a unique directed path to $\mathcal{Q}$. Assign weight $\pi(g) := \prod_{(i \rightarrow j) \in g} p_{ij}$ to each graph.

We also define $\mathcal{G}(i \nearrow \mathcal{Q})$ as the set of graphs with exactly $r - 1$ edges from non-target vertices, no directed cycles, and no directed path from i to $\mathcal{Q}$. Equivalently, $\mathcal{G}(i \nearrow \mathcal{Q})$ consists of all $\mathcal{Q}$ -graphs from which a single edge on the (unique) path from i to $\mathcal{Q}$ has been removed.

With the above notation, we have an explicit formula for the expected hitting time, as follows: for $i \in \mathcal{V} \setminus \mathcal{Q}$, the expected hitting time satisfies the following formula:

$$h_i = \frac{\sum_{g \in \mathcal{G}(i \nearrow \mathcal{Q})} \pi(g)}{\sum_{g \in \mathcal{G}(\mathcal{Q})} \pi(g)}. \quad (5.8)$$

The formula and derivation below are not new; see [Freidlin and Wentzell \(1998b, Chap. 6, Lem. 3.3\)](#) for the general statement. We include them here to shed light on the key ingredients of the analysis.

We now verify this formula i.e., that the right-hand side satisfies the equations (5.7).

Proof of the claim. Let $N_i := \sum_{g \in \mathcal{G}(i \nearrow \mathcal{Q})} \pi(g)$ and $D := \sum_{g \in \mathcal{G}(\mathcal{Q})} \pi(g)$. We verify that $h_i = N_i/D$ satisfies the hitting-time equations. The case $i \in \mathcal{Q}$ is immediate since $N_i = 0$ by convention. For $i \notin \mathcal{Q}$, using $\sum_j p_{ij} = 1$, the equation $h_i = 1 + \sum_j p_{ij} h_j$ is equivalent to

$$\sum_{j \neq i} p_{ij} h_j = 1 + \sum_{j \neq i} p_{ij} h_j. \quad (5.9)$$

Substituting $h = N/D$ and multiplying through by D , (5.9) becomes

$$\sum_{j \neq i} \sum_{g \in \mathcal{G}(i \nearrow \mathcal{Q})} p_{ij} \pi(g) = \sum_{g \in \mathcal{G}(\mathcal{Q})} \pi(g) + \sum_{j \neq i} \sum_{g \in \mathcal{G}(j \nearrow \mathcal{Q})} p_{ij} \pi(g). \quad (5.10)$$

We establish (5.10) by a bijective argument on the terms of the left-hand side. Consider a pair (j, g) with $j \neq i$ and $g \in \mathcal{G}(j \nearrow \mathcal{Q})$, contributing weight $p_{ij} \cdot \pi(g)$ to the left-hand side. Two cases arise, according to whether j belongs to the

subtree of g rooted at i —that is, the set of vertices from which a directed path in g leads to i .

Case 1: j does not belong to the subtree of g rooted at i . Then $g \cup \{(i \rightarrow j)\} \in \mathcal{G}(\mathcal{Q})$: adding $(i \rightarrow j)$ creates no directed cycle (since j has no directed path to i in g), and i can now reach $\mathcal{Q}$. The map $(j, g) \mapsto g \cup \{(i \rightarrow j)\}$ is a bijection with $\mathcal{G}(\mathcal{Q})$.² The weights match: $\pi(g \cup \{(i \rightarrow j)\}) = p_{ij} \cdot \pi(g)$.

Case 2: j belongs to the subtree of g rooted at i , so there exists a directed path $j \rightarrow \dots \rightarrow i$ in g . Since g has no directed path from i to $\mathcal{Q}$, j also has no directed path to $\mathcal{Q}$ in g . By definition, $g \in \mathcal{G}(j \not\rightarrow \mathcal{Q})$, so the pair (j, g) contributes $p_{ij} \cdot \pi(g)$ to the second sum on the right-hand side.

The two cases are mutually exclusive and exhaustive, so summing over all pairs (j, g) establishes (5.10). ■

We now apply (5.8) to the accelerated induced chain. By Proposition 5.1, its transition probabilities satisfy $Q_{\mathcal{C}}^{K\delta}(x, \mathcal{C}_j^\delta) \approx \exp(-\mathcal{B}_{ij}/\eta)$ for $x \in \mathcal{C}_i^\delta$. Substituting and letting η be small,

$$\begin{aligned} h_i &\sim \frac{\sum_{g \in \mathcal{G}(i \not\rightarrow \mathcal{Q})} \exp\left(-\frac{1}{\eta} \sum_{(j \rightarrow k) \in g} \mathcal{B}_{jk}\right)}{\sum_{g \in \mathcal{G}(\mathcal{Q})} \exp\left(-\frac{1}{\eta} \sum_{(j \rightarrow k) \in g} \mathcal{B}_{jk}\right)} \\ &\sim \exp\left(\frac{\min_{g \in \mathcal{G}(i \not\rightarrow \mathcal{Q})} \sum_{(j \rightarrow k) \in g} \mathcal{B}_{jk} - \min_{g \in \mathcal{G}(\mathcal{Q})} \sum_{(j \rightarrow k) \in g} \mathcal{B}_{jk}}{\eta}\right), \end{aligned} \quad (5.11)$$

since, for small η , each exponential sum concentrates on its minimum-weight term. This leads us to introduce $E(\mathcal{Q} | i)$ as the difference of minimum tree weights, defined precisely below.

Definition 5.1. For $i \in \mathcal{V} \setminus \mathcal{Q}$, define:

$$E(\mathcal{Q}) := \min_{g \in \mathcal{G}(\mathcal{Q})} \sum_{(k \rightarrow l) \in g} \mathcal{B}_{k,l}, \quad (5.12)$$

$$E(i \not\rightarrow \mathcal{Q}) := \min_{g \in \mathcal{G}(i \not\rightarrow \mathcal{Q})} \sum_{(k \rightarrow l) \in g} \mathcal{B}_{k,l}, \quad (5.13)$$

$$E(\mathcal{Q} | i) := E(\mathcal{Q}) - E(i \not\rightarrow \mathcal{Q}). \quad (5.14)$$

The quantity $E(\mathcal{Q})$ represents the minimum aggregate cost to route all non-target components toward $\mathcal{Q}$; it can thus be seen as an overall measure of how easy it is to reach $\mathcal{Q}$. $E(i \not\rightarrow \mathcal{Q})$ is the same cost with i excluded from the routing requirement. Their difference $E(\mathcal{Q} | i)$ captures the additional cost of routing i to $\mathcal{Q}$, i.e., the hardness of reaching $\mathcal{Q}$ starting from i ; it plays a central role in our estimates of the global convergence time.

5.2.2 Refined setup

To study the hitting time to an arbitrary component, we need to refine the setup of the previous chapters. The main difference from Chapter 4 is that we now

² Its inverse sends $h \in \mathcal{G}(\mathcal{Q})$ to $(h(i), h \setminus \{(i \rightarrow h(i))\})$, where $h(i)$ is the unique vertex that i points to in h .

care about *when* SGD reaches the global minimizers of f , not just the long-run distribution.

Our main goal is thus to characterize the hitting time of SGD to the global minimizers of f . To this end, we partition the critical components of f into two groups: the targets (global minimizers) and the non-targets (everything else that SGD must escape). Note that our framework also allows us to analyze the hitting time to any subset of critical components, not just the global minimizers, see [Appendix B.1](#).

Definition 5.2 (Choice of components). Denote by $\mathcal{Q}_1, \dots, \mathcal{Q}_{N_{\text{targ}}}$ the connected components of $\text{crit}(f)$ that are global minimizers i.e.,

$$\arg \min f = \bigcup_{j=1}^{N_{\text{targ}}} \mathcal{Q}_j, \quad (5.15)$$

and define the *target set* $\mathcal{Q} := \{1, \dots, N_{\text{targ}}\}$. Denote by $\mathcal{K}_1, \dots, \mathcal{K}_K$ the remaining connected components of $\text{crit}(f)$. Define

$$(\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}) := (\mathcal{K}_1, \dots, \mathcal{K}_K, \mathcal{Q}_1, \dots, \mathcal{Q}_{N_{\text{targ}}}) \quad (5.16)$$

with $N_{\text{crit}} = K + N_{\text{targ}}$.

This partition is the natural one for the hitting-time problem. For the invariant measure analysis of [Definition 4.1](#), all critical components played symmetric roles; here, the targets $\mathcal{Q}_1, \dots, \mathcal{Q}_{N_{\text{targ}}}$ are treated as absorbing — meaning we stop tracking the iterates once they enter a neighborhood of any $\mathcal{Q}_j$, $j \in \mathcal{Q}$, as a convenient analytical device rather than a feature of the algorithm itself.

Echoing the results of [Chapter 4](#), the relevant notion of transition cost becomes the quasi-potential *conditioned* on avoiding $\mathcal{Q} = \bigcup_{j \in \mathcal{Q}} \mathcal{Q}_j$ at intermediate integer times. Concretely, for $x, x' \in \mathbb{R}^d$, define

$$\begin{aligned} \mathcal{B}(x, x') &:= \inf \{ \mathcal{S}_T(\gamma) : T \geq 1, \gamma \in \mathcal{C}_r(T), \gamma_0 = x, \gamma_T = x' \} \\ &= \inf \{ \mathcal{S}_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 = x, \xi_{N-1} = x' \}, \end{aligned} \quad (5.17)$$

where $\mathcal{C}_r(T)$ is the set of continuous paths of duration T whose values at integer times $1, \dots, T-1$ avoid the target, and $\mathcal{D}_r(N)$ is the analogous set of discrete paths of length N . The two infima coincide by definition of ρ (see [\(3.43\)](#)).

This differs from the transition costs of [§ 4.1.1](#): there, paths were free to pass through any component, since the invariant measure analysis assigned no preferred role to any one component. Here, conditioning on avoiding $\mathcal{Q}$ ensures that $\mathcal{B}(x, x')$ governs *first* passages to the global minimizers: a path that hits $\mathcal{Q}$ before reaching x' is excluded, as it would have already converged. Thus $\mathcal{B}(x, x')$ represents the cost of the least-action path joining x to x' while avoiding the global minimizers in between. The transition costs between components are defined in the same way as before:

$$\mathcal{B}_{i,j} := \inf \{ \mathcal{B}(x, x') : x \in \mathcal{C}_i, x' \in \mathcal{C}_j \}. \quad (5.18)$$

Like before, we assume for simplicity that the transition costs between components are finite (see [Assumption 3.10](#) for details).

Assumption 5.3 (Same as [Assumption 3.10](#)). For any $i \in \{1, \dots, K\}$ and $j \in \{1, \dots, N_{\text{crit}}\}$, $\mathcal{B}_{i,j} < +\infty$.

As for the results on the invariant measure [Chapter 4](#), our analysis relies on approximating SGD by a finite-state Markov chain that tracks which component neighborhood the process is in, see [Definition 3.5](#). To be able to control the hitting time, we need to refine the construction of the induced chain. Put simply, the induced chain is now obtained by keeping only the iterates of $(x_n^\eta)_n$ that enter some $\mathcal{C}_i^\delta$ and discarding all iterates in between (see [Figure 18](#)).

Definition 5.4 (Induced chain). Consider $\mathcal{C}_i^\delta$ δ -neighborhoods of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, and let $\mathcal{C}^\delta := \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$. Define recursively the stopping times $(\sigma_n^\eta)_{n \geq 0}$ and $(\tau_n^\eta)_{n \geq 0}$ by:

- $\sigma_0^\eta := 0, \tau_0^\eta := 0$.
- For $n \geq 0$, if $x_{\tau_n^\eta}^\eta \in \mathcal{C}_{i_n}^\delta$ for some i_n :

$$\sigma_{n+1}^\eta := \begin{cases} \inf\{k \geq \tau_n^\eta : x_k^\eta \notin \mathcal{C}_{i_n}^\delta\} & \text{if } \mathcal{C}_{i_n} \in \{\mathcal{C}_1, \dots, \mathcal{C}_K\}, \\ \tau_n^\eta & \text{otherwise,} \end{cases} \quad (5.19)$$

$$\tau_{n+1}^\eta := \inf\{k \geq \sigma_{n+1}^\eta : x_k^\eta \in \mathcal{C}^\delta\}. \quad (5.20)$$

The induced chain is $z_n := x_{\tau_n^\eta}^\eta, n \geq 0$, with transition probabilities $\mathbb{Q}_{\mathcal{C}^\delta}$.

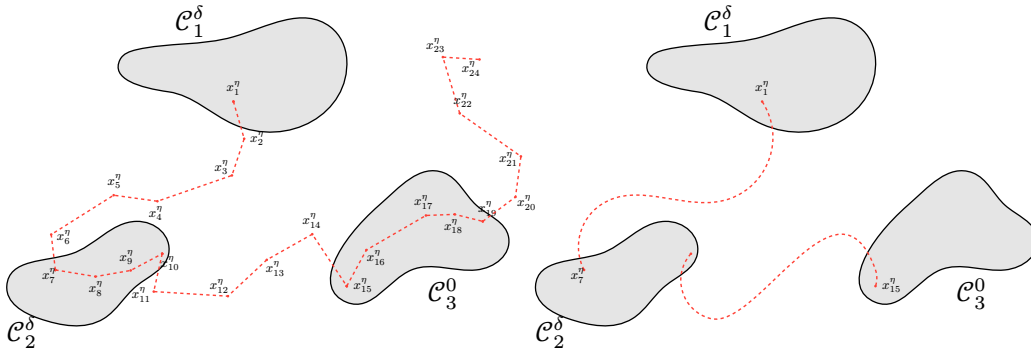


Figure 18: Illustration of the construction of the induced chain: left, the full trajectory of SGD; right, the same trajectory after suppressing the iterates that are not part of the induced chain.

Note that $z_n \in \mathcal{C}^\delta$ for all $n \geq 1$, and that once the chain enters $\bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$, it remains there forever. This definition is more refined than [Definition 3.5](#): the stopping times $(\sigma_n^\eta, \tau_n^\eta)_n$ track each visit to $\mathcal{C}^\delta$ individually, which is needed for precise time estimates.

In an analogous way to [Proposition 3.20](#), we will work with an accelerated version of the induced chain by subsampling every K steps (where K is the number of non-target components [Definition 5.2](#)). Taking the K -step power ensures that, after one step of the accelerated chain, the process has had enough time to potentially visit every non-target component at least once.

Definition 5.5 (Accelerated induced chain). Given the induced chain $(z_n)_{n \geq 0}$, define the *accelerated induced chain* by

$$z_n^K := z_{nK}. \quad (5.21)$$

Its transition probabilities are denoted by $\mathbb{Q}_{\mathcal{C}^\delta}^K$.

The following proposition shows that the transition probabilities of the accelerated induced chain satisfy the same exponential estimates as [Proposition 3.20](#), with the quasi-potential $\mathcal{B}$ now conditioned on avoiding $\mathcal{Q}$. The proof is given in [Appendix A.3.5](#).

To simplify the statements, for the rest of this chapter, we fix a precision level $\varepsilon > 0$, and consider sufficiently small $\mathcal{C}_i^\delta$ δ -neighborhoods of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$ and sufficiently small $\eta > 0$.

Proposition 5.1. *For any $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$,*

$$\forall x \in \mathcal{C}_i^\delta, \quad \exp\left(-\frac{\mathcal{B}_{i,j} + \varepsilon}{\eta}\right) \leq \mathbf{Q}^K_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) \leq \exp\left(-\frac{\mathcal{B}_{i,j} - \varepsilon}{\eta}\right). \quad (5.22)$$

The next two results bound the *time* to make one transition, showing that it remains essentially bounded. In particular, transition times do not contribute to the leading exponential term in the global hitting time: what dominates the global convergence time of SGD is the hitting time of the induced chain, not the time it takes to make the transitions between components. More precisely, the expected time to make one transition of the accelerated induced chain is at most $e^{\varepsilon/\eta}$, which is negligible compared to the exponential hitting time of the induced chain, which scales as $e^{E(\mathcal{Q}|i)/\eta}$ by [\(5.11\)](#). Hence the dominant factor is the transition probability structure of the induced chain, not the speed of individual transitions.

Proposition 5.2. *Fix a compact set $\mathcal{X}_0 \subset \mathbb{R}^d$, which does not intersect the target $\mathcal{Q} = \bigcup_{j \in \mathcal{Q}} \mathcal{Q}_j$. Uniformly over all initializations $x \in \mathcal{X}_0$, the expected duration of one transition ([Definition 5.4](#)) satisfies:*

$$1 \leq \mathbb{E}_x[\tau_1^\eta] \leq e^{\varepsilon/\eta}. \quad (5.23)$$

Note the same bound also applies to the duration of one transition of the accelerated induced chain by the strong Markov property. This bound relies partially on [Proposition 3.11](#), which guarantees that SGD returns to a neighborhood of the critical set in sub-exponential time.

5.2.3 Hitting time for accelerated SGD

We now combine the three ingredients established above to obtain an exponential two-sided bound on the expected hitting time of the accelerated induced chain. [Proposition 5.1](#) provides exponential bounds on the transition probabilities; [\(5.8\)](#) converts these into the combinatorial tree formula for the expected hitting time involving $E(\mathcal{Q}|i)$ ([Definition 5.1](#)); and [Proposition 5.2](#) controls the duration of each individual transition, ensuring it is sub-exponential. These results will be then transferred to the actual SGD process in the next section [§ 5.2.4](#).

We first state the result when the initial state lies in a neighborhood $\mathcal{C}_i^\delta$ of a non-target component, and then extend to arbitrary initial points in [Theorem 5.4](#).

In words, when SGD starts in $\mathcal{C}_i^\delta$, the expected hitting time of the accelerated induced chain ([Definition 5.5](#)) satisfies, for small η ,

$$\mathbb{E}_x[\tau_Q^\eta] \sim \exp\left(\frac{E(\mathcal{Q}|i)}{\eta}\right) \quad (5.24)$$

The energy $E(\mathcal{Q} | i)$ therefore plays the same role for convergence time as the quasi-potential $\mathcal{B}$ plays for transition probabilities: it is the exponent governing how long SGD must wait before reaching $\mathcal{Q}$.

Theorem 5.3. Fix $i \notin \mathcal{Q}$ the index of a non-target component. For any initialization $x \in \mathcal{C}_i^\delta$,

$$\exp\left(\frac{E(\mathcal{Q} | i) - \varepsilon}{\eta}\right) \leq \mathbb{E}_x[\tau_{\mathcal{Q}}^\eta] \leq \exp\left(\frac{E(\mathcal{Q} | i) + \varepsilon}{\eta}\right). \quad (5.25)$$

The proof combines (5.8) with Proposition 5.1 and Proposition 5.2: Proposition 5.1 gives the exponential approximation for the transition probabilities of the accelerated induced chain, from which the bounds on h_i follow by concentration of exponential sums as $\eta \rightarrow 0$.

It remains to extend Theorem 5.3 to arbitrary initial points. For this, given an initial point x_0 , we need to characterize the probability that SGD first reaches $\mathcal{C}_i$ before any other. It involves the following quantity: for any $x \in \mathbb{R}^d$ and $i \in \mathcal{V} \setminus \mathcal{Q}$, define

$$\mathcal{B}(x, \mathcal{C}_i) := \inf\{\mathcal{B}(x, x') : x' \in \mathcal{C}_i\}. \quad (5.26)$$

Then, the probability that SGD first reaches $\mathcal{C}_i$ is approximately

$$\mathbb{P}_{x_0}(\text{SGD first reaches } \mathcal{C}_i^\delta) \approx \exp\left(-\frac{\mathcal{B}(x_0, \mathcal{C}_i)}{\eta}\right). \quad (5.27)$$

Given an arbitrary initial point $x_0 \in \mathbb{R}^d$, we now decompose the expected hitting time via the law of total expectation, conditioning on which component SGD first reaches:

$$\begin{aligned} \mathbb{E}_{x_0}[\tau_{\mathcal{Q}}^\eta] &= \sum_{i \in \mathcal{V} \setminus \mathcal{Q}} \mathbb{P}_{x_0}(\text{SGD first reaches } \mathcal{C}_i^\delta) \cdot \mathbb{E}_{x_0}[\tau_{\mathcal{Q}}^\eta \mid \text{SGD first reaches } \mathcal{C}_i^\delta] \\ &\approx \sum_{i \in \mathcal{V} \setminus \mathcal{Q}} \exp\left(-\frac{\mathcal{B}(x_0, \mathcal{C}_i)}{\eta}\right) \cdot \exp\left(\frac{E(\mathcal{Q} | i)}{\eta}\right). \end{aligned} \quad (5.28)$$

The first factor $\mathbb{P}_{x_0}(\text{SGD first reaches } \mathcal{C}_i^\delta)$ is the probability that x_0 is first attracted to $\mathcal{C}_i$; the second factor is the conditional expected hitting time from $\mathcal{C}_i^\delta$, which equals $\exp(E(\mathcal{Q} | i)/\eta)$ by Theorem 5.3. For small η , the sum concentrates on its maximum-weight term, giving

$$\mathbb{E}_{x_0}[\tau_{\mathcal{Q}}^\eta] \sim \exp\left(\frac{\max_{i \in \mathcal{V} \setminus \mathcal{Q}} (E(\mathcal{Q} | i) - \mathcal{B}(x_0, \mathcal{C}_i))}{\eta}\right). \quad (5.29)$$

This motivates the following definition, with the addition of the positive part $[\cdot]_+$ to ensure non-negativity.

Definition 5.6. For $x \in \mathbb{R}^d$, define

$$E(\mathcal{Q} | x) := \left[\max_{i \in \mathcal{V} \setminus \mathcal{Q}} (E(\mathcal{Q} | i) - \mathcal{B}(x, \mathcal{C}_i)) \right]_+, \quad (5.30)$$

with the convention that it equals $E(\mathcal{Q} | i)$ if $x \in \mathcal{C}_i^\delta$, and 0 if $x \in \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$.

The maximum over i accounts for the fact that x may be attracted to different components before the main transition chain begins, as seen in (5.28).

Theorem 5.4. For any initialization $x_0 \in \mathbb{R}^d$,

$$\exp\left(\frac{E(\mathcal{Q} | x_0) - \varepsilon}{\eta}\right) \leq \mathbb{E}_{x_0}[\tau_{\mathcal{Q}}^{\eta}] \leq \exp\left(\frac{E(\mathcal{Q} | x_0) + \varepsilon}{\eta}\right). \quad (5.31)$$

In other words, the expected hitting time from an arbitrary initial point scales as

$$\mathbb{E}_{x_0}[\tau_{\mathcal{Q}}^{\eta}] \sim \exp\left(\frac{E(\mathcal{Q} | x_0)}{\eta}\right). \quad (5.32)$$

The energy $E(\mathcal{Q} | x_0)$ thus provides a single scalar summary of the global convergence difficulty from any starting point: it accounts for both the cost of reaching the critical region and the subsequent inter-component transition costs.

5.2.4 Global hitting time of SGD

We now translate the hitting-time bounds for the accelerated process to the original SGD sequence (SGD). Define the *global hitting time* of SGD:

$$\tau := \inf\left\{n \geq 0 : x_n \in \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^{\delta}\right\}. \quad (5.33)$$

The connection to the accelerated process is immediate: by construction, $\tau \leq \tau_{\mathcal{Q}}^{\eta} \cdot \lceil \eta^{-1} \rceil$, and $\lceil \eta^{-1} \rceil \leq e^{\varepsilon/\eta}$ for small enough η . Hence the upper bound transfers directly from Theorem 5.4.

Theorem 5.5 (Upper bound). For any initialization $x_0 \in \mathbb{R}^d$,

$$\mathbb{E}_{x_0}[\tau] \leq \exp\left(\frac{E(\mathcal{Q} | x_0) + \varepsilon}{\eta}\right). \quad (5.34)$$

The proof follows directly from Theorem 5.4 and $\tau \leq \tau_{\mathcal{Q}}^{\eta} \cdot \lceil \eta^{-1} \rceil \leq \tau_{\mathcal{Q}}^{\eta} \cdot e^{\varepsilon/\eta}$. The lower bound requires more care. The upper bound used that if subsampled SGD reaches $\mathcal{Q}$, then SGD also reaches $\mathcal{Q}$. For the lower bound, we need the converse — that if SGD reaches $\mathcal{Q}$, then subsampled SGD also reaches $\mathcal{Q}$. For this, we need to ensure that once SGD reaches a neighborhood of $\mathcal{Q}$, it remains there for a long time with high probability, so that the accelerated process also touches $\mathcal{Q}$. This requires a local stability assumption around each target component.

For this, we thus *attracting strength* of $\mathcal{Q}$ as the maximal value of the product μR^2 , where μ and R are such that, for any $x \in \mathbb{R}^d$ with $\text{dist}(x, \mathcal{Q}) \leq R$,

$$\langle \nabla f(x), x - x^* \rangle \geq \mu \|x - x^*\|^2 \quad (5.35)$$

with $x^* \in \mathcal{Q}$ denoting a projection of x onto $\mathcal{Q}$. In this way, (5.35) can be seen as a “setwise” second-order optimality condition for the global minimum of f . Intuitively, the larger the attracting strength of $\mathcal{Q}$, the sharper it is. For the lower bound to hold, we will thus require that the attracting strength of $\mathcal{Q}$ be large enough. We should note here that the requirement on the attracting strength for the lower bound is purely technical, see Appendix B.3.3 for details. For instance, given an objective function f , it is possible to ensure that its global minimum has large enough attracting strength by locally modifying f around its global minima without changing the overall landscape.

The next result leverages this local stability to establish a lower bound on the hitting time of SGD which complements the upper bound of [Theorem 5.5](#).

Theorem 5.6. *Fix $i \notin \mathcal{Q}$ the index of a non-target component. Assuming that the attracting strength of $\mathcal{Q}$ is large enough, for any initialization $x_0 \in \mathcal{C}_i^\delta$,*

$$\mathbb{E}_{x_0}[\tau] \geq \exp\left(\frac{E(\mathcal{Q} | i) - \varepsilon}{\eta}\right). \quad (5.36)$$

The proof uses the lower bound of [Theorem 5.3](#) for the accelerated induced chain, combined with a local stability lemma for SGD near each target component. Indeed, assume that SGD reaches a neighborhood of $\mathcal{Q}$ around some $\mathcal{Q}_j$ at time τ . Then, having large attracting strength ensures that, with high probability, SGD remains within distance at most R from $\mathcal{Q}_j$ forever. Since the transition times are bounded [Proposition 5.2](#), this implies that the accelerated process also reaches $\mathcal{Q}$ at some time $\tau_{\mathcal{Q}}^\eta \sim \tau/\eta$. Since [Theorem 5.3](#) lower bounds $\mathbb{E}_x[\tau_{\mathcal{Q}}^\eta]$ by $\exp(E(\mathcal{Q} | i)/\eta)$, this implies that $\mathbb{E}_x[\tau]$ is also lower bounded by $\exp(E(\mathcal{Q} | i)/\eta)$. The requirement that the attracting strength be large enough ensures that this scenario happens with high enough probability not to influence the expectation of τ . The full proof is in [Appendix B.3.3](#). Combining the upper and lower bounds [Theorems 5.5](#) and [5.6](#), we obtain the characterization of the global convergence time of SGD.

Theorem 5.7 (Global convergence time). *For any initialization $x_0 \in \mathbb{R}^d$:*

$$\exp\left(\frac{E(\mathcal{Q} | x_0) - \varepsilon}{\eta}\right) \leq \mathbb{E}_{x_0}[\tau] \leq \exp\left(\frac{E(\mathcal{Q} | x_0) + \varepsilon}{\eta}\right), \quad (5.37)$$

provided the attracting strength of $\mathcal{Q}$ is large enough for the lower bound to hold.

[Theorem 5.7](#) characterizes the global convergence time of (SGD) up to multiplicative errors:

$$\mathbb{E}_{x_0}[\tau] \sim \exp\left(\frac{E(\mathcal{Q} | x_0)}{\eta}\right). \quad (5.38)$$

The bounds scale exponentially with the inverse step-size η , and show that this convergence time is determined by $E(\mathcal{Q} | x_0)$. In this sense, $E(\mathcal{Q} | x_0)$ encodes the hardness of the non-convex optimization landscape and captures the fact that the difficulty of global convergence stems from the presence of spurious local minima, i.e., locally minimizing components that are not globally minimizing. In particular, $E(\mathcal{Q} | x_0)$ vanishes if and only if there are no spurious local minima, in which case τ becomes subexponential.

Note also that the global convergence time is not driven by the shortest, *least costly* chain of transitions from x_0 to $\arg \min f$, but by the longest, *most costly* one: even if the direct path to $\mathcal{Q}$ is easy, the algorithm may fall into a deep spurious trap whose escape time dominates everything else. The next section refines these results when the initialization lies in a specific basin of attraction.

5.2.5 Basin-dependent global hitting time of SGD

The bound in [Theorem 5.7](#) depends on the initialization x_0 only through $E(\mathcal{Q} | x_0)$, which involves the quasi-potential $\mathcal{B}(x_0, \mathcal{C}_i)$ from x_0 to each non-target component. When x_0 lies in the basin of attraction $\text{Attr}(\mathcal{C}_i)$ of some minimizing component ([Definition 4.3](#)) $\mathcal{C}_i$ for the gradient flow of f ([Definition 3.1](#)), this cost is strictly positive for all other components, so the dominant term simplifies to

$E(\mathcal{Q} | i)$. Formally, the *basin of attraction* $\text{Attr}(\mathcal{C}_i)$ of a component $\mathcal{C}_i \subset \text{crit}(f)$ is defined as the set of points $x \in \mathbb{R}^d$ such that the gradient flow trajectory starting at x converges to $\mathcal{C}_i$.

The key ingredient is the following lemma, which ensures that exiting a basin of a minimizing component requires strictly positive action cost.

Lemma 5.8. *For any minimizing component $\mathcal{C}_i \subset \text{crit}(f)$ and any $x \in \text{Attr}(\mathcal{C}_i)$:*

$$\mathcal{B}\left(x, \mathbb{R}^d \setminus \text{Attr}(\mathcal{C}_i)\right) > 0. \quad (5.39)$$

The proof relies on the observation that any low-cost path starting in $\text{Attr}(\mathcal{C}_i)$ approximately follows the gradient flow, which converges to $\mathcal{C}_i$. In other words, escaping $\text{Attr}(\mathcal{C}_i)$ requires climbing uphill away from $\mathcal{C}_i$; since $\mathcal{C}_i$ is a local minimizer, any such escape necessarily passes through a region where f is strictly higher, incurring positive action cost. Full details are in [Appendix B.3.4](#).

Now, [Lemma 5.8](#) implies that if x_0 is in $\text{Attr}(\mathcal{C}_i)$ for some minimizing $\mathcal{C}_i$, then with high probability the process first reaches a neighborhood of $\mathcal{C}_i$ before any other component. Conditioned on this event, [Theorem 5.3](#) applies directly and yields a sharp bound in terms of $E(\mathcal{Q} | i)$ alone for the hitting time of the accelerated induced chain.

Theorem 5.10 (Basin-dependent accelerated hitting time). *Suppose $x_0 \in \text{Attr}(\mathcal{C}_i)$ for some minimizing $i \in \{1, \dots, N_{\text{crit}}\}$. There is an event H , and a constant $A > 0$ such that*

$$\mathbb{P}_{x_0}(H) \geq 1 - e^{-A/\eta} \quad (5.40)$$

and

$$\exp\left(\frac{E(\mathcal{Q} | i) - \varepsilon}{\eta}\right) \leq \mathbb{E}_{x_0}[\tau_Q^\eta H] \leq \exp\left(\frac{E(\mathcal{Q} | i) + \varepsilon}{\eta}\right). \quad (5.41)$$

Similarly to how [Theorem 5.7](#) was obtained from [Theorem 5.4](#), this result for accelerated hitting time can be transferred to the actual hitting time of SGD by leveraging the local stability of SGD near each target component.

Theorem 5.10. *Suppose $x_0 \in \text{Attr}(\mathcal{C}_i)$ for some minimizing $i \in \{1, \dots, N_{\text{crit}}\}$. If the attracting strength of $\mathcal{Q}$ is large enough, then there is an event H , and a constant $A > 0$ such that:*

$$\mathbb{P}_{x_0}(H) \geq 1 - e^{-A/\eta} \quad (5.42)$$

and

$$\exp\left(\frac{E(\mathcal{Q} | i) - \varepsilon}{\eta}\right) \leq \mathbb{E}_{x_0}[\tau H] \leq \exp\left(\frac{E(\mathcal{Q} | i) + \varepsilon}{\eta}\right). \quad (5.43)$$

Taken together, [Theorem 5.10](#) shows that, conditioned on a high-probability event, the global convergence time of (SGD) starting from any $x_0 \in \text{Attr}(\mathcal{C}_i)$ is determined by $E(\mathcal{Q} | i)$, the energy of $\mathcal{Q}$ relative to $\mathcal{C}_i$, the minimizing component that attracts x_0 under the gradient flow of f :

$$\mathbb{E}_{x_0}[\tau] \sim \exp\left(\frac{E(\mathcal{Q} | i)}{\eta}\right). \quad (5.44)$$

Intuitively, all initial points in the basin of $\mathcal{C}_i$ move to a small neighborhood of $\mathcal{C}_i$ with overwhelmingly high probability and, because this transition takes relatively little time, the global convergence time to $\mathcal{Q}$ is approximately equal to the convergence time from $\mathcal{C}_i$ to $\mathcal{Q}$. Under standard non-degeneracy assumptions, Lebesgue-almost every point lies in the basin of some minimizing component.

However, the argument in this section does not yield a simplification of [Theorem 5.7](#) for arbitrary initial points. The next sections quantify how $E(\mathcal{Q} | i)$ depends on the geometry of the loss landscape.

5.3 EXTENSIONS AND REFINEMENTS

In this section, we present extensions and refinements of our main results. First, in [§ 5.3.1](#), we examine the case where there are no spurious local minima, showing that the energy $E(\mathcal{Q} | x_0)$ vanishes for every initialization. In [§ 5.3.2](#), we explore further the link between the energy $E(\mathcal{Q} | x_0)$ and the geometry of the loss landscape, showing that the energy depends on the depth of local minima. In [§ 5.3.3](#), we provide numerical illustrations of the main results. In [§ 5.3.4](#), we revisit [§ 4.2.3](#): leveraging the hitting-time estimates of [§ 5.2](#), we derive non-asymptotic versions of the results on occupation measures.

5.3.1 No spurious local minima

To start understanding the properties of the energy $E(\mathcal{Q} | x_0)$, we start by examining the extreme case: when is $E(\mathcal{Q} | x_0)$ equal to zero? This actually corresponds to the case where there are no spurious local minima, as the following proposition shows.

Proposition 5.11. *The following properties are equivalent:*

- (i) For all $x \in \mathbb{R}^d$, $E(\mathcal{Q} | x) = 0$.
- (ii) For all $i \in \{1, \dots, N_{\text{crit}}\}$, $E(\mathcal{Q} | i) = 0$.
- (iii) No non-target component C_i , $i \in \{1, \dots, N_{\text{crit}}\} \setminus \mathcal{Q}$, is locally minimizing.

In other words, the energy $E(\mathcal{Q} | x_0)$ vanishes for every initialization if and only if there are no spurious local minima. In that case, [Theorem 5.5](#) implies that $\mathbb{E}[\tau] \leq e^{\varepsilon/\eta}$ for any $\varepsilon > 0$, so the global convergence time is subexponential in $1/\eta$. In the context of neural network training, our results offer a first way to connect geometric properties of the loss landscape to sharp estimates of the convergence time of (SGD). Indeed, under certain conditions, neural networks have no spurious local minima ([Nguyen and Hein, 2017](#), [Nguyen et al., 2019a](#)), in which case [Proposition 5.11](#) guarantees *subexponential* convergence to the global minimum of f .

In this case, our results complement the existing literature on the avoidance of saddle points, e.g., ([Ge et al., 2015b](#)). These results guarantee that SGD hits an ε -neighborhood of a global minimizer within a time polynomial in $1/\varepsilon$, though under more stringent assumptions, including that all minimizers have positive-definite Hessian and all saddle points are strict.

5.3.2 Influence of the loss landscape

In this section, we explain how the energy $E(\mathcal{Q} | i)$ of [Theorem 5.10](#), which controls the global convergence time, depends on the noise, the depths of the spurious minima and their relative positions with respect to each other. We sketch the key elements and we refer to [Appendix B.4](#) for the full details.

In particular, for simplicity, we assume here that the subGaussian parameter $\sigma_\infty^2(f(x))$ ([Assumption 3.3\(c\)](#)) is bounded by a constant $\bar{\sigma}^2$ for all x in some large enough compact set.

Assumption 5.7. The cumulant-generating function of U is bounded as

$$\log \mathbb{E} \left[e^{\langle p, U(x, \omega) \rangle} \right] \leq \frac{1}{2} \bar{\sigma}^2 \|p\|^2 \quad \text{for all } p \in \mathbb{R}^d, \quad (5.45)$$

for all x in some large enough compact set.

The first step is to derive upper and lower bounds on the transition costs $\mathcal{B}_{i,j}$. From [Assumption 5.7](#), we have a lower-bound on $\mathcal{B}_{i,j}$ for any i, j : we have $\mathcal{B}_{i,j} \geq \underline{B}_{ij}$ where

$$\underline{B}_{ij} := \inf_{T, \gamma} \left\{ \sup_{s < t} \frac{2(f(\gamma(t)) - f(\gamma(s)))}{\bar{\sigma}^2} : \gamma(0) \in \mathcal{C}_i, \gamma(T) \in \mathcal{C}_j \right\}. \quad (5.46)$$

This means that the transition cost is at least equal to the maximal upward jump in the objective function that cannot be avoided when going from $\mathcal{C}_i$ to $\mathcal{C}_j$. To obtain the upper-bound, we need an extra assumption on the noise.

EXTRA NOISE ASSUMPTION. While we only assumed a bound on the magnitude of the gradient noise ([Assumption 3.3\(c\)](#)), we now require in addition a bound on the “minimal level of noise”, through the following assumption.

Assumption 5.8. The Lagrangian is bounded as

$$\mathcal{L}(x, v) \leq \frac{\|v + \nabla f(x)\|^2}{2\underline{\sigma}^2} \quad \forall v \text{ s.t. } \|v\| \leq 2\|\nabla f(x)\|$$

for all x in some large enough compact set.

This assumption is satisfied by general types of noise (including finite-sum models) given a lower-bound on the variance of $Z(x; \omega)$ and a condition on the support; see [Appendix B.4.1](#). In particular, if $Z(x; \omega)$ follows a (truncated) Gaussian distribution with a variance σ^2 (as in [§ 4.2.1](#)) then [Assumption 5.8](#) is satisfied with $\underline{\sigma}^2 = (1 - \varepsilon)\sigma^2$, where ε depends on the truncation level. In that case, the constant $\bar{\sigma}^2$ of [Assumption 3.3\(c\)](#) and [Assumption 5.7](#) can be taken as $\bar{\sigma}^2 = (1 + \varepsilon)\sigma^2$ (see [Appendix B.4.1](#)). Note then that the ratio $\underline{\sigma}^2/\bar{\sigma}^2$ is close to 1, which will matter in the next theorem.

TRANSITION BETWEEN BASINS. With [Assumption 5.8](#), we obtain quantitative upper-bounds on the transition costs between two neighboring components. More precisely, if $\mathcal{C}_i$ and $\mathcal{C}_j$ are such that their basins of attraction intersect, then the cost of transitioning from $\mathcal{C}_i$ to $\mathcal{C}_j$ is bounded by

$$\mathcal{B}_{i,j} \leq \frac{2(f_{i,j} - f_i)}{\underline{\sigma}^2} \quad (5.47)$$

where $f_{i,j}$ is the minimum of f over that intersection and f_i is the value of f on $\mathcal{C}_i$.

With these bounds, we recover immediately the example of [§ 5.1](#). When the closure of the basin of $\mathcal{C}_i$ intersects $\mathcal{C}_j$ itself, $f_{i,j}$ is simply the value of f on $\mathcal{C}_j$ and (5.47) simplifies to the same expression as in (5.2). The equality in (5.2) is thus obtained by combining (5.46) and (5.47) in the Gaussian case.

GRAPH STRUCTURE. From the bounds of $\mathcal{B}_{i,j}$, we can get bounds on $E(\mathcal{Q} | i)$ as follows. We restrict the transition graph of [§ 5.2](#) to neighboring components:

we consider $\mathcal{G}'$ with vertices $\mathcal{V}$ and edges $j \rightsquigarrow k$ if the closures of the basins of $\mathcal{C}_j$ and $\mathcal{C}_k$ intersect. Given a path $\mathcal{P}_j = j \rightsquigarrow j_1 \rightsquigarrow \dots \rightsquigarrow j_m$ in $\mathcal{G}'$ that ends in $\mathcal{Q}$, we define its cost as

$$\text{cost}(\mathcal{P}_j) := \max_{n=0, \dots, m-1} \max \{f_{j_n, j_{n+1}} - f_{j_n}, f_{j_n, j_{n+1}} - f_j\}.$$

This cost involves the values of the objective function along the path from j to $\mathcal{Q}$ and captures the maximum depth of a minima encountered on the path: indeed, $f_{j_n, j_{n+1}} - f_{j_n}$ represents the jump in the loss function when going from $\mathcal{C}_{j_n}$ to $\mathcal{C}_{j_{n+1}}$ while $f_{j_n, j_{n+1}} - f_j$ represents the total jump when going from $\mathcal{C}_j$ to $\mathcal{C}_{j_{n+1}}$.

RESULT AND DISCUSSION. With the above quantities, we can establish that the energy $E(\mathcal{Q} | i)$, that governs the global convergence time of [Theorem 5.10](#), is bounded as follows.

Theorem 5.11. Under [Assumption 5.8](#), for $i \in \mathcal{V} \setminus \mathcal{Q}$,

$$E(\mathcal{Q} | i) \leq \max_{j: \underline{B}_{ij} \leq r} \min_{\mathcal{P}_j} \frac{2 \text{cost}(\mathcal{P}_j)}{\underline{\sigma}^2} + \mathcal{O}\left(1 - \frac{\sigma^2}{\underline{\sigma}^2}\right) \quad (5.48)$$

for some $r > 0$ that depends on the graph structure $\mathcal{G}'$.

This means that $E(\mathcal{Q} | i)$ is bounded by the maximum depth of the minimizers that SGD must go through in order to reach $\mathcal{Q}$. This involves all the paths in $\mathcal{G}'$ that start at a component close to i , as measured by $\underline{B}_{ij}$. By forgetting the graph, we can obtain a simpler bound that highlights how the convergence time depends on the depth of the spurious minima:

Corollary 5.13. Under [Assumption 5.8](#), for any $i \in \mathcal{V} \setminus \mathcal{Q}$,

$$E(\mathcal{Q} | i) \leq \frac{2(\max_{\text{crit}(f)} f - \min_{\text{spurious}(f)} f)}{\underline{\sigma}^2} + \mathcal{O}\left(1 - \frac{\sigma^2}{\underline{\sigma}^2}\right), \quad (5.49)$$

where $\text{spurious}(f)$ denotes the set of spurious local minima of f .

Though this bound loses its dependency on the initialization, it shows that the convergence time depends on (i) the height of the highest barrier to the global minimum and (ii) the depth of the deepest spurious local minimum.

Consider, for instance, the example of [§ 5.1](#): the bound of [Theorem 5.11](#) is attained for $p_5 \rightsquigarrow p_3 \rightsquigarrow p_1$ and we recover the formula of [\(5.4\)](#), which also corresponds to [Corollary 5.13](#) in that case. Since [\(5.4\)](#) is exact, this shows that [Theorem 5.11](#) and [Corollary 5.13](#) are tight in that example.

In this case, the convergence time depends exponentially on the depths of the minima. However, different noise models would lead to qualitatively different results: Gaussian noise with variance scaling linearly with the objective value ([Mori et al., 2022](#)) would lead to a polynomial dependence on the depths of the minima. Our framework in [Appendix B.4](#) encompasses this case. Informally, as for [\(4.39\)](#) in [§ 4.2.1](#), in the case of (truncated) Gaussian noise $\mathcal{N}(0, \sigma^2(f(x))I)$ with $\sigma^2(f(x)) = \sigma_1^2(f(x) + \sigma_0^2)$ for some $\sigma_0, \sigma_1 > 0$, the bound of the analogue of [Corollary 5.13](#) would read:

$$E(\mathcal{Q} | i) \leq \frac{2(\max_{\text{crit}(f)} \log(f + \sigma_0^2) - \min_{\text{spurious}(f)} \log(f + \sigma_0^2))}{\sigma_1^2} \quad (5.50)$$

$$+ \mathcal{O}\left(1 - \frac{\sigma^2}{\bar{\sigma}^2}\right).$$

Combined with [Theorem 5.10](#), this implies that in that case the global convergence time depends polynomially on the depths of the minima and not exponentially.

Importantly, in the context of neural network training, our results offer a way to connect geometric properties of the loss landscape to sharp estimates of the convergence time of (SGD). Indeed, under certain conditions, neural networks have no spurious local minima ([Nguyen and Hein, 2017](#), [Nguyen et al., 2019a](#)), in which case [Theorem 5.7](#) guarantees *subexponential* convergence to the global minimum of f . More to the point, even when spurious minima of bounded depth are present ([Nguyen et al., 2021](#)), The combination of [Theorem 5.10](#) and [Theorem 5.11](#) provide a way to translate these depth bounds directly into convergence time estimates for (SGD).

5.3.3 Numerical illustrations

We provide some more numerical experiments as validation of the estimates provided by [Theorem 5.10](#) and [Corollary 5.13](#).

For concreteness, we consider the following multimodal test functions:

(a) *The Styblinski–Tang function:*

$$f(x_1, x_2) = \frac{1}{2} \sum_{i=1}^d (x_i^4 - 16x_i^2 + 5x_i) \quad (5.51a)$$

(b) *The Himmelblau function:*

$$f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2 \quad (5.51b)$$

In [Figures 19](#) and [20](#), we plot the landscape of f for each of the above cases, as well as the runtime statistics of SGD in each case. As in [Figure 17](#), we considered different values of the algorithm’s learning rate η , and for each value of η , we performed 500 runs of SGD with Gaussian noise ($\sigma = 50$). Then, for each “particle” of SGD, we recorded the number of iterations required to reach the vicinity of p_1 (within $\delta = 10^{-2}$) from the same, fixed initialization. The resulting boxplots are displayed in [Figures 19](#) and [20](#) in log-inverse scale, with a point density overlay to illustrate the full empirical distribution of τ for every η .

In both cases, we plotted the theoretical prediction of [Theorem 5.10](#) (labeled “Theory”), against a linear regression plot of the observed runtime data (labeled “Data”).³ For the Styblinski–Tang function, we observe a situation similar to the three-hump-camel example of [Figure 17](#): Both theory and experiment indicate that the global convergence time τ of SGD scales exponentially in the inverse of the algorithm’s step-size η , and there is excellent agreement between theory and experiment ($R^2 = 0.973$ and $R^2 = 0.989$ respectively). By contrast, in the Himmelblau case, runtimes are significantly slower, and the slope of the theoretical estimator is 0: This is because the Himmelblau function does not have any spurious minima, so τ enters the subexponential regime (see [§ 5.3.1](#)). This is confirmed by the fact that the algorithm’s runtime is several orders of

³ In more detail, we plotted a linear fit with slope given by [Corollary 5.13](#).

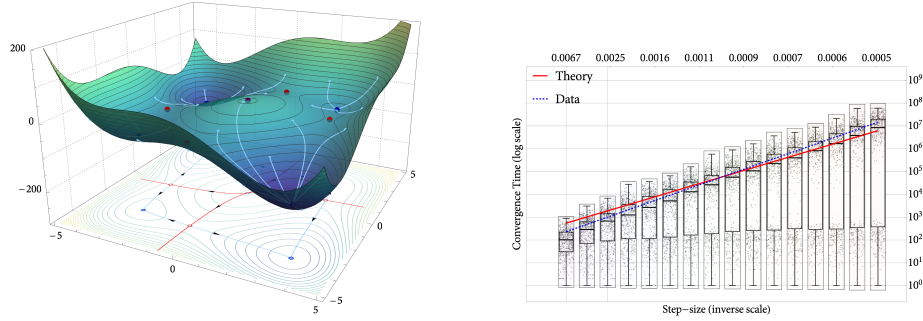


Figure 19: Visualization and hitting time statistics for the Styblinski–Tang function (5.51a). The loss landscape (left) is superimposed on the stable and unstable manifolds of the function’s critical points, which define the edges of its transition graph (the arrows indicate the direction of the gradient flow, and hence the transitions which come at no cost).

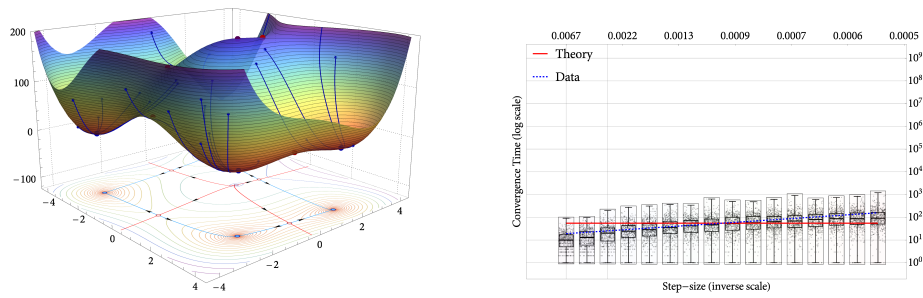


Figure 20: Visualization and hitting time statistics for the Himmelblau function (5.51b). The loss landscape (left) is superimposed on the stable and unstable manifolds of the function’s critical points, which define the edges of its transition graph (the arrows indicate the direction of the gradient flow, and hence the transitions which come at no cost).

magnitude smaller, with strong agreement between our theoretical estimate and the numerically measured values ($R^2 = 0.913$ and $R^2 = 0.976$ respectively).

With the three-hump camel function of § 5.1, we also illustrate the case of noise variance scaling linearly with f , as in (5.50): $\sigma(f(x)) = 25^2 f(x) + 1$. In this regime, the global convergence time depends polynomially (rather than exponentially) on the depths of the spurious minima but still scales exponentially in the inverse of the step-size η .

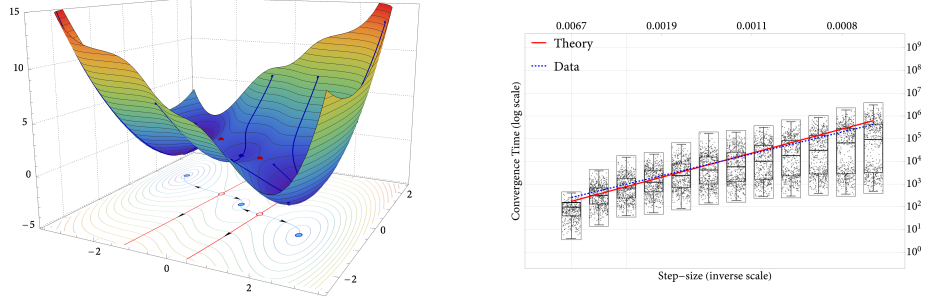


Figure 21: Visualization and hitting time statistics for the three-hump camel function of § 5.1 with noise variance scaling linearly with $f(x)$. The loss landscape (left) is superimposed on the stable and unstable manifolds of the function’s critical points, as in Figure 16.

5.3.4 Non-asymptotic occupation measures

We now revisit § 4.2.3. These theorems are asymptotic in n : they describe the long-run behavior of μ_n but do not quantify how large n must be. We now provide the non-asymptotic counterparts, using the hitting-time bounds of this chapter. The key idea is elementary: the occupation measure of a set $\mathcal{B}$ decomposes as a transient phase (before SGD first reaches $\mathcal{B}$, contributing at most $\tau_{\mathcal{B}}/n$ to $\mu_n(\mathcal{B}^c)$) and a stationary phase (after arrival, where the invariant-measure estimates apply). The hitting-time upper bound from Theorem 5.5 then converts the exponential bound on τ into an explicit $1/n$ correction on top of the asymptotic occupation-measure guarantee.

We first give a non-asymptotic version of Theorem 4.6*, which guarantees that SGD spends most of its time near the critical components. Let $\mathcal{C}^\delta := \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$ be a union of small neighborhoods of all critical components.

Theorem 5.14. Fix an initialization $x_0 \in \mathbb{R}^d$. There exists $c > 0$ such that, for all $n \geq 1$:

$$\mu_n(\mathcal{C}^\delta) \geq 1 - \frac{1}{n} \exp\left(\frac{E(\mathcal{C}^\delta | x_0) + \varepsilon}{\eta}\right) - e^{-c/\eta}. \quad (5.52)$$

The bound (5.52) has two terms. The exponential term $e^{-c/\eta}$ is the same as in the asymptotic Theorem 4.6*: it reflects the probability that the invariant measure assigns to the complement of $\mathcal{C}^\delta$. The new $1/n$ term captures the transient phase before SGD first enters $\mathcal{C}^\delta$; its numerator $\exp((E(\mathcal{C}^\delta | x_0) + \varepsilon)/\eta)$ is the hitting-time upper bound from Theorem 5.5. In particular, for $n \gg \exp(E(\mathcal{C}^\delta | x_0)/\eta)$, the $1/n$ term becomes negligible and the result reduces to the asymptotic guarantee.

We now give non-asymptotic Gibbs bounds for each individual component. The lower bound involves the hitting time to reach $\mathcal{C}_i^\delta$ from x_0 ; the upper bound involves the hitting time to reach the target set $\mathcal{Q}$ of global minimizers.

Theorem 5.15. Fix an initialization $x_0 \in \mathbb{R}^d$ and a component index $i \in \{1, \dots, N_{\text{crit}}\}$. Then, for all $n \geq 1$

$$\mu_n(\mathcal{C}_i^\delta) \geq \left(1 - \frac{1}{n} \exp\left(\frac{E(\{i\} | x_0) + \varepsilon}{\eta}\right)\right) \cdot \exp\left(-\frac{E_i - \min_j E_j + \varepsilon}{\eta}\right), \quad (5.53)$$

$$\mu_n(\mathcal{C}_i^\delta) \leq \left(1 + \frac{2}{n}\right) \cdot \exp\left(-\frac{E_i - \min_j E_j - \varepsilon}{\eta}\right). \quad (5.54)$$

The lower bound (5.53) says that the occupation measure of $\mathcal{C}_i^\delta$ is at least the asymptotic Gibbs weight $\exp(-(E_i - \min E + \varepsilon)/\eta)$, reduced by a factor accounting for the transient before SGD first reaches $\mathcal{C}_i^\delta$. The upper bound (5.54) inflates the asymptotic Gibbs weight by a factor that reflects the time before the process reaches any global minimizer.

In both cases the $1/n$ correction vanishes once n exceeds the relevant hitting time, recovering the asymptotic Theorem 4.3*.

When x_0 lies in the basin of a specific component, we could replace $E(\cdot | x_0)$ by $E(\cdot | i)$, which depends only on the basin and not on the specific starting point within it, similarly to Theorem 5.10.

We give the non-asymptotic version of Theorem 4.7*, which is the most practically relevant: it quantifies how long SGD must run to ensure that its occupation measure is concentrated on the ground state.

Theorem 5.16. Fix an initialization $x_0 \in \mathbb{R}^d$. There exists $c > 0$ such that, for all $n \geq 1$

$$\mu_n(\mathcal{C}_0^\delta) \geq 1 - \frac{1}{n} \exp\left(\frac{E(\mathcal{Q} | x_0) + \varepsilon}{\eta}\right) - e^{-c/\eta}. \quad (5.55)$$

The bound (5.55) has the same two-term structure as (5.52). The exponential term $e^{-c/\eta}$ is the same as in the asymptotic Theorem 4.7*: it reflects the probability that the invariant measure assigns to the complement of $\mathcal{C}_0^\delta$. The new $1/n$ term captures the transient phase before SGD first enters $\mathcal{C}_0^\delta$; its numerator $\exp((E(\mathcal{Q} | x_0) + \varepsilon)/\eta)$ is the hitting-time upper bound from Theorem 5.5. In particular, for $n \gg \exp(E(\mathcal{Q} | x_0)/\eta)$, the $1/n$ term becomes negligible and the result reduces to the asymptotic guarantee.

Finally, we state a basin-dependent version of this result.

Theorem 5.17. Suppose $x_0 \in \text{Attr}(C_i)$ for some minimizing $i \in \{1, \dots, N_{\text{crit}}\}$. There exist $A > 0$, $c > 0$ and an event H with $\mathbb{P}_{x_0}(H) \geq 1 - e^{-A/\eta}$ such that for all $n \geq 1$:

$$\mathbb{E}_{x_0} \left[\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}\{x_k \in \mathcal{C}_0^\delta\} \mid H \right] \geq 1 - \frac{1}{n} \exp\left(\frac{E(\mathcal{Q} | i) + \varepsilon}{\eta}\right) - e^{-c/\eta}. \quad (5.56)$$

Theorem 5.17 sharpens Theorem 5.16 when the initialization lies in a specific basin: the transient term now depends only on the component index i via $E(\mathcal{Q} | i)$, rather than on the full initial condition x_0 . This mirrors the basin-dependent hitting-time result of Theorem 5.10 and can now be combined with the explicit bounds of § 5.3.2 to yield non-asymptotic occupation-measure guarantees that depend explicitly on the geometry of the loss landscape.

Part II

DISTRIBUTIONALLY ROBUST AND CONSTRAINED MIN-MAX OPTIMIZATION

WASSERSTEIN DISTRIBUTIONALLY ROBUST OPTIMIZATION: EXACT GENERALIZATION GUARANTEES AND REGULARIZATION

6.1	Introduction	95
6.1.1	Generalization and (Wasserstein) Distributionally Robust Models	95
6.1.2	Contributions and related works	97
6.2	Setup and Assumptions	97
6.2.1	Wasserstein Distributionally Robust risk functions	98
6.2.2	Blanket assumptions	98
6.3	Main results and discussions	99
6.3.1	Exact generalization guarantees for WDRO models	99
6.3.2	Idea of the proofs	101
6.4	Regularized WDRO	103
6.4.1	Regularizing the WDRO problem	103
6.4.2	Dual formulation of the regularized WDRO problem	104
6.4.3	Approximation of the regularized WDRO problem	105
6.4.4	Generalization guarantees for the regularized WDRO problem	106
6.4.5	Examples	107
6.5	Conclusion	109

This chapter is primarily based on [Azizian et al. \(2023c\)](#). In addition, § 6.4 synthesizes material from [Azizian et al. \(2023a\)](#), which is reproduced in full in [Appendix C](#). In particular, compared to § 6.4, [Azizian et al. \(2023a\)](#) covers two additional aspects: (i) arbitrary convex regularizers (beyond the entropic case), and (ii) unbounded samples spaces (beyond the compact setting).

6.1 INTRODUCTION

6.1.1 Generalization and (Wasserstein) Distributionally Robust Models

We consider the fundamental question of generalization of machine learning models. Let us denote by f_θ the loss induced by a model parametrized by θ for some uncertain variable ξ (typically a data point). When ξ follows some distribution P , seeking the best parameter θ writes as minimizing the expected loss

$$\min_{\theta \in \Theta} \mathbb{E}_{\xi \sim P}[f_\theta(\xi)]. \quad (6.1)$$

We usually do not have a direct knowledge of P but rather we have access to samples $(\xi_i)_{i=1}^n$ independently drawn from P . The empirical risk minimization

approach then consists in minimizing the expected loss over the associated empirical distribution $\widehat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$ (as a proxy for the expected loss over P), i.e.,

$$\min_{\theta \in \Theta} \mathbb{E}_{\xi \sim \widehat{P}_n} [f_\theta(\xi)] \left(= \frac{1}{n} \sum_{i=1}^n f_\theta(\xi_i) \right). \quad (6.2)$$

Classical statistical learning theory ensures that, with high probability, $\mathbb{E}_P[f_\theta]$ is close to $\mathbb{E}_{\widehat{P}_n}[f_\theta]$ up to $\mathcal{O}(1/\sqrt{n})$ error terms, see e.g., the monographs [Boucheron et al.](#), [Wainwright \(2019\)](#).

A practical drawback of empirical risk minimization is that it can lead to over-confident decisions (when $\mathbb{E}_{\widehat{P}_n}[f_\theta] < \mathbb{E}_P[f_\theta]$, the real loss can be higher than the empirical one ([Esfahani and Kuhn, 2018](#))). In addition, this approach is also sensitive to distribution shifts between training and application. To overcome these drawbacks, an approach gaining momentum in machine learning is *distributionally robust* optimization, which consists in minimizing the *worst expectation* of the loss when the distribution lives in a neighborhood of $\widehat{P}_n$:

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{U}(\widehat{P}_n)} \mathbb{E}_{\xi \sim Q} [f_\theta(\xi)], \quad (6.3)$$

where the inner sup is thus taken over Q in the neighborhood $\mathcal{U}(\widehat{P}_n)$ of $\widehat{P}_n$ in the space of probability distributions. Popular choices of distribution neighborhoods are based on the KL divergence ([Laguel et al., 2020](#), [Levy et al., 2020](#)), kernel tools ([Zhu et al., 2021a](#), [Staib and Jegelka, 2019](#), [Zhu et al., 2021b](#)), moments ([Delage and Ye, 2010](#), [Goh and Sim, 2010](#)), or Wasserstein distance ([Shafieezadeh Abadeh et al., 2015](#), [Esfahani and Kuhn, 2018](#)). If $P \in \mathcal{U}(\widehat{P}_n)$, distributionally robust models can benefit from direct generalization guarantees as,

$$\sup_{Q \in \mathcal{U}(\widehat{P}_n)} \mathbb{E}_{\xi \sim Q} [f_\theta(\xi)] \geq \mathbb{E}_{\xi \sim P} [f_\theta(\xi)]. \quad (6.4)$$

Thus, for well-chosen neighborhoods $\mathcal{U}(\widehat{P}_n)$, distributionally robust objectives are able to provide *exact upper-bounds* on the expected loss over distribution P , i.e., the true risk.

WDRO problems correspond to (6.3) with

$$\mathcal{U}(\widehat{P}_n) = \left\{ Q \in \mathcal{P}(\Xi) : W(\widehat{P}_n, Q) \leq \rho \right\}, \quad (6.5)$$

where $W(\widehat{P}_n, Q)$ denotes the Wasserstein distance between $\widehat{P}_n$ and Q and $\rho > 0$ controls the required level of robustness around $\widehat{P}_n$. As a natural metric to compare discrete and absolutely continuous probability distributions, the Wasserstein distance has attracted a lot of interest in both machine learning ([Shafieezadeh Abadeh et al., 2015](#), [Sinha et al., 2018](#), [Shafieezadeh-Abadeh et al., 2019b](#), [Li et al., 2020a](#), [Kwon et al., 2020](#)) and operation research ([Zhao and Guan, 2018](#), [Arrigo et al., 2022](#)) communities; see e.g., the review articles [Blanchet et al. \(2021a\)](#), [Kuhn et al. \(2019\)](#).

WDRO benefits from out-of-the-box generalization guarantees in the form of (6.4) since it inherits the concentration properties of the Wasserstein distance ([Esfahani and Kuhn, 2018](#)). More precisely, under mild assumptions on P , ([Fournier and Guillin, 2015](#)) establishes that $W(\widehat{P}_n, P) \leq \rho$ with high probability

as soon as $\rho \sim 1/n^{1/d}$ where d denotes the dimension of the samples space. Thus, a major issue is the prescribed radius ρ suffers from the curse of the dimensionality: when d is large, ρ decreases slowly as the number of samples n increases. This contrasts with other distributionally robust optimization ambiguity sets, such as maximum mean discrepancy (MMD) (Staib and Jegelka, 2019, Zeng and Lam, 2022), where the radius scales as $1/\sqrt{n}$. Moreover, the existing scaling for WDRO is overly conservative for WDRO objectives since recent works (Blanchet et al., 2022a, Blanchet and Shapiro, 2023) prove that a radius behaving as $1/\sqrt{n}$ is asymptotically optimal. The main difference with (Esfahani and Kuhn, 2018) is that they — and us — consider the WDRO objective as a whole, instead of proceeding in two steps: first considering the Wasserstein distance independently and invoking concentration results on the Wasserstein distance and then plugging this result in the WDRO problem.

6.1.2 Contributions and related works

In this chapter, we show that WDRO provides exact upper-bounds on the true risk with high probability. More precisely, we prove non-asymptotic generalization bounds of the form of (6.4), that hold for general classes of functions, and that only require ρ to scale as $1/\sqrt{n}$ and not $1/n^{1/d}$. To do so, we construct an interval for the radius ρ for which it is both sufficiently large so that we can go from the empirical to the true estimator (i.e., at least of the order of $1/\sqrt{n}$) and sufficiently small so that the robust problem does not become degenerate (i.e., smaller than some critical radius, that we introduce as an explicit constant). Our results imply proving concentration results on Wasserstein Distributionally Robust objectives that are of independent interest.

This work is part of a rich and recent line of research about theoretical guarantees on WDRO for machine learning. One of this first results, Lee and Raginsky (2018), provides generalization guarantees, for a general class of models and a fixed ρ , that, however, become degenerate as the radius goes to zero. In the particular case of linear models, WDRO models admit an explicit form that allows Shafieezadeh-Abadeh et al. (2019b), Chen and Paschalidis (2018) to provide generalization guarantees (6.4) with the radius scaling as $1/\sqrt{n}$. The case of general classes of models, possibly non-linear, is more intricate. Sinha et al. (2018) showed that a modified version of (6.4) holds at the price of non-negligible error terms. Gao (2022), An and Gao (2021) made another step towards broad generalization guarantees for WDRO but with error terms that vanish only when ρ goes to zero.

In contrast, our analysis provides exact generalization guarantees in the form (6.4) without additional error terms, that hold for general classes of functions and allow for a non-vanishing uncertainty radius to cover for distribution shifts at testing. Moreover, our guarantees also carry over to the recently introduced regularized versions of WDRO (Wang et al., 2021, Azizian et al., 2023a), whose statistical properties have not been studied yet.

This paper is organized as follows. In § 6.2, we introduce notations and our blanket assumptions. In § 6.3, we present our main results, an idea of proof, and discussions. The complete proofs are deferred to the appendix.

6.2 SETUP AND ASSUMPTIONS

In this section, we formalize our setting and introduce Wasserstein Distributionally Robust risks.

6.2.1 Wasserstein Distributionally Robust risk functions

In this paper, we consider as a samples space Ξ a subset of $\mathbb{R}^d$ equipped with the Euclidean norm $\|\cdot\|$. We rely on Wasserstein distances of order 2, in line with the seminal work [Blanchet et al. \(2022a\)](#) on generalization of WDRO. This distance is defined for two distributions Q, Q' in the set of probability distributions on Ξ , denoted by $\mathcal{P}(\Xi)$, as

$$W_2(Q, Q') := \left(\inf_{\pi \in \mathcal{P}(\Xi \times \Xi), \pi_1=Q, \pi_2=Q'} \mathbb{E}_{(\xi, \zeta) \sim \pi} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \right)^{1/2}, \quad (6.6)$$

where $\mathcal{P}(\Xi \times \Xi)$ is the set of probability distributions in the product space $\Xi \times \Xi$, and π_1 (resp. π_2) denotes the first (resp. second) marginal of π .

We denote by $f : \Xi \rightarrow \mathbb{R}$ the loss function of some model over the sample space. The model may depend on some parameter θ , that we drop for now to lighten the notations; instead, we consider a class of functions $\mathcal{F}$ encompassing our various models and losses of interest (we come back to classes of parametric models of the form $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ in [§ 6.4.5](#)).

We define the empirical Wasserstein Distributionally Robust risk $\widehat{\mathcal{R}}_{\rho^2}(f)$ centered on $\widehat{P}_n$ and similarly the true robust risk $\mathcal{R}_{\rho^2}(f)$ centered on P as

$$\widehat{\mathcal{R}}_{\rho^2}(f) := \sup_{\substack{Q \in \mathcal{P}(\Xi) \\ W_2^2(\widehat{P}_n, Q) \leq \rho^2}} \mathbb{E}_{\xi \sim Q} [f(\xi)] \quad \text{and} \quad \mathcal{R}_{\rho^2}(f) := \sup_{\substack{Q \in \mathcal{P}(\Xi) \\ W_2^2(P, Q) \leq \rho^2}} \mathbb{E}_{\xi \sim Q} [f(\xi)]. \quad (6.7)$$

Note that $\widehat{\mathcal{R}}_{\rho^2}(f)$, which is based on the empirical distribution $\widehat{P}_n$, is a computable proxy for the true robust risk $\mathcal{R}_{\rho^2}(f)$. Note also that the true robust risk $\mathcal{R}_{\rho^2}(f)$ immediately upper-bounds the true (non-robust) risk $\mathbb{E}_{\xi \sim P} [f(\xi)]$ and also upper-bounds $\mathbb{E}_{\xi \sim Q} [f(\xi)]$ for neighboring distributions Q that correspond to distributions shifts of magnitude smaller than ρ in Wasserstein distance.

We introduce regularized versions of these risks in [§ 6.4](#), where entropic regularization of WDRO is studied in detail.

6.2.2 Blanket assumptions

Our analysis is carried under the following set of assumptions that will be in place throughout the paper. First, we assume that the sample space $\Xi \subset \mathbb{R}^d$ is convex and compact, which is in line with previous work, e.g., ([Lee and Raginsky, 2018](#), [An and Gao, 2021](#)).

Assumption 6.1 (On the set Ξ). The sample space Ξ is a compact convex subset of $\mathbb{R}^d$.

Second, we require the class of loss functions $\mathcal{F}$ to be sufficiently regular. In particular, we assume that they have Lipschitz continuous gradients.

Assumption 6.2 (On the function class). The functions of $\mathcal{F}$ are twice differentiable, uniformly bounded, and their derivatives are uniformly bounded and uniformly Lipschitz.

Finally, we assume that $\widehat{P}_n$ is made of i.i.d. samples of P and that P is supported on the interior of Ξ (which can be done without loss of generality by slightly enlarging Ξ if needed).

Assumption 6.3 (On the distributions). $\widehat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$ where $\xi_1, \dots, \xi_n$ are i.i.d. samples of P . We further assume that there is some $R > 0$ such that P satisfies $\text{supp } P + \overline{B}(0, R) \subset \Xi$.

6.3 MAIN RESULTS AND DISCUSSIONS

The main results of our paper establish that the empirical robust risk provide high probability bounds, of the form of (6.4), on the true risk. Generalization guarantees for the regularized case are deferred to § 6.4. In § 6.3.2, we provide the common outline for the proofs of these results, the proofs themselves being provided in the appendix. Finally, in § 6.4.5, we detail some examples.

6.3.1 Exact generalization guarantees for WDRO models

In this section, we require the two following additional assumptions on the function class. The first assumption is common in the WDRO litterature, see e.g., Blanchet et al. (2022a), Blanchet and Shapiro (2023), Gao (2022), An and Gao (2021).

Assumption 6.4. The quantity $\inf_{f \in \mathcal{F}} \mathbb{E}_P [\|\nabla f\|^2]$ is positive.

The second assumption we consider in this section makes use of the notation $d(\xi, A)$, for a set $A \subset \Xi$ and a point $\xi \in \Xi$, to denote the distance between ξ and A , i.e., $d(\xi, A) = \inf_{\zeta \in A} \|\xi - \zeta\|$.

Assumption 6.5.

1. For any $R > 0$, there exists $\Delta > 0$ such that,

$$\forall f \in \mathcal{F}, \forall \zeta \in \Xi, \quad d(\zeta, \arg \max f) \geq R \implies f(\zeta) - \max f \leq -\Delta. \quad (6.8)$$

2. The following growth condition holds: there exist $\mu > 0$ and $L > 0$ such that, for all $f \in \mathcal{F}$, $\xi \in \Xi$ and ξ^* a projection of ξ on $\arg \max f$, i.e., $\xi^* \in \arg \min_{\arg \max f} \|\xi - \cdot\|$,

$$f(\xi^*) \geq f(\xi) + \frac{\mu}{2} \|\xi - \xi^*\|^2 - \frac{L}{6} \|\xi - \xi^*\|^3. \quad (6.9)$$

The first item of this assumption has a natural interpretation: we show in Lemma D.1.7, that it is equivalent to the relative compactness of the function space $\mathcal{F}$ w.r.t. to the distance

$$D(f, g) := \|f - g\|_\infty + D_H(\arg \max f, \arg \max g), \quad (6.10)$$

where D_H denotes the (Hausdorff) distance between sets and $\|f\|_\infty := \sup_{\xi \in \Xi} |f(\xi)|$ is the infinity norm. The last one is a structural assumption on the functions $\mathcal{F}$ that is new in our context but is actually very close the so-called parametric Morse-Bott condition, introduced in of bilevel optimization (Arbel and Mairal, 2022), see Appendix D.1.5.

We now state our main generalization result for WDRO risks.

Theorem 6.1. Under [Assumptions 6.4](#) and [6.5](#), there is an explicit constant ρ_c depending only on $\mathcal{F}$ and $\mathbb{P}$ such that for any $\delta \in (0, 1)$ and $n \geq 1$, if

$$\mathcal{O}\left(\sqrt{\frac{1 + \log 1/\delta}{n}}\right) \leq \rho \leq \frac{\rho_c}{2} - \mathcal{O}\left(\sqrt{\frac{1 + \log 1/\delta}{n}}\right) \quad (6.11)$$

then, there is $\rho_n = \mathcal{O}\left(\sqrt{\frac{1 + \log 1/\delta}{n}}\right)$ such that, with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathbb{E}_{\xi \sim Q}[f(\xi)] \quad \text{for all } Q \text{ such that } W_2^2(\mathbb{P}, Q) \leq \rho(\rho - \rho_n). \quad (6.12)$$

In particular, with probability $1 - \delta$, we have

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathbb{E}_{\xi \sim \mathbb{P}}[f(\xi)]. \quad (6.13)$$

The second part of the result, [\(6.13\)](#), is an *exact* generalization bound: it is an actual upper-bound on the true risk $\mathbb{E}_{\xi \sim \mathbb{P}}[f(\xi)]$, that we cannot access in general, through a quantity that we can actually compute with $\widehat{\mathbb{P}}_n$. The first part of the result, [\(6.12\)](#) gives us insight into the robustness guarantees offered by the WDRO risk. Indeed, it tells us that, when ρ is greater than the minimal radius $\rho_n \propto 1/\sqrt{n}$ by some margin, the empirical robust risk $\widehat{\mathcal{R}}_{\rho^2}(f)$ is an upper-bound on the loss even with some perturbations of the true distribution. Hence, as long as ρ is large enough, the WDRO objective enables us to guarantee the performance of our model even in the event of a distribution shift at testing time. In other words, the empirical robust risk is an *exact* upper-bound on the true robust risk $\mathcal{R}_{\rho(\rho - \rho_n)}(f)$ with a reduced radius.

The range of admissible radiuses is described by [\(6.11\)](#). The lower-bound, roughly proportional to $1/\sqrt{n}$, is optimal, following the results of [Blanchet et al. \(2022a\)](#). The upper-bound, almost independent of n , depends on a constant ρ_c , that we call *critical radius* and that has an interesting interpretation, that we formalize in the following remark. Note, finally, that, the big-O notation in this theorem has a slightly stronger meaning than the usual one, being non-asymptotic in n and δ : e.g., $\rho_n = \mathcal{O}\left(\sqrt{\frac{1 + \log 1/\delta}{n}}\right)$ means that $\exists C > 0$ such that $\rho_n \leq C\sqrt{\frac{1 + \log 1/\delta}{n}}$ for all $\delta \in (0, 1)$ and $n \geq 1$.

Remark 7 (Interpretation of critical radius). The critical radius ρ_c , appearing in [\(6.11\)](#), is defined by

$$\rho_c^2 := \inf_{f \in \mathcal{F}} \mathbb{E}_{\xi \sim \mathbb{P}} \left[\frac{1}{2} d^2(\xi, \arg \max f) \right]. \quad (6.14)$$

It can be interpreted as the threshold at which the WDRO problem w.r.t. $\mathbb{P}$ starts becoming degenerate. Indeed, when $\rho^2 \geq \mathbb{E}_{\xi \sim \mathbb{P}} \left[\frac{1}{2} d^2(\xi, \arg \max f) \right]$ for some $f \in \mathcal{F}$ that we fix, the distribution Q given by the second marginal of the transport plan π defined by,

$$\pi(d\xi, d\zeta) := \mathbb{P}(d\xi) \delta_{\zeta^*(\xi)}(d\zeta) \quad \text{where} \quad \zeta^*(\xi) \in \arg \min_{\zeta \in \arg \max f} d^2(\xi, \zeta), \quad (6.15)$$

satisfies

$$W_2^2(P, Q) \leq \mathbb{E}_{(\xi, \zeta) \sim \pi} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] = \mathbb{E}_{\xi \sim P} \left[\frac{1}{2} d^2(\xi, \arg \max f) \right] \leq \rho^2. \quad (6.16)$$

As a consequence, the robust problem is equal to

$$\mathcal{R}_{\rho^2}(f) = \sup_{\substack{Q \in \mathcal{P}(\Xi) \\ W_2^2(P, Q) \leq \rho^2}} \mathbb{E}_{\xi \sim Q} [f(\xi)] = \max_{\xi \in \Xi} f(\xi). \quad (6.17)$$

Thus, when the radius exceeds ρ_c , there is some f such that the robust problem becomes degenerate as it does not depend on P nor ρ anymore.

Finally, note that we can obtain the same generalization guarantee as [Theorem 6.1](#) without [Assumption 6.5](#) at the expense of losing the above interpretation on the condition on the radius. More precisely, we have the following result.

Theorem 6.2. *Let [Assumption 6.4](#) hold. For any $\delta \in (0, 1)$ and $n \geq 1$, if ρ satisfies [\(6.11\)](#), and if, in addition, it is smaller than a positive constant which depends only on $P, \mathcal{F}$ and Ξ , then both conclusions of [Theorem 6.1](#) hold.*

This theorem can be compared to existing results, and in particular with [Gao \(2022\)](#), [An and Gao \(2021\)](#). These two papers provide generalization bounds for WDRO under a similar assumption on $\mathcal{F}$ and a weakened version of [Assumption 6.3](#). However, these generalization bounds involve extra error terms, that require ρ to be vanishing. In comparison, with a similar set of assumptions, [Theorem 6.2](#) improves on these two issues, by allowing ρ not to vanish as $n \rightarrow \infty$ and by providing the exact upper-bound [\(6.13\)](#). Allowing non-vanishing radiuses is an attractive feature of our results that enables us to cover distribution shifts.

6.3.2 Idea of the proofs

In this section, we present the main ideas of the proofs of [Theorem 6.1](#) and [Theorem 6.2](#). The full proofs are detailed in [Appendix D](#); we point to relevant sections along the discussion. A similar argument applies to the regularized case; see [§ 6.4.4](#).

First, we recall the duality results for WDRO that play a crucial role in our analysis. Second, we present a rough sketch of the proofs. Finally, we provide a refinement of our results that is a by-product of our analysis.

DUALITY IN WDRO. Duality has been a central tool in both the theoretical analyses and computational schemes of WDRO from the onset ([Shafieezadeh Abadeh et al., 2015](#), [Esfahani and Kuhn, 2018](#)). The expressions of the dual of WDRO problems ([Gao and Kleywegt, 2022](#), [Blanchet and Murthy, 2019](#)) can be written with the following dual generator function ϕ defined as

$$\phi(f, \xi, \lambda) := \sup_{\zeta \in \Xi} \left\{ f(\zeta) - \frac{\lambda}{2} \|\xi - \zeta\|^2 \right\}, \quad (6.18)$$

where λ is the dual variable associated to the Wasserstein constraint in [\(6.7\)](#). The regularized counterpart of this dual generator is derived in [§ 6.4.2](#). Then,

under the blanket assumptions of § 6.2.2, the existing strong duality results yield that the empirical robust risk writes

$$\widehat{\mathcal{R}}_{\rho^2}(f) = \inf_{\lambda \geq 0} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda)] , \quad (6.19)$$

and, similarly, the true robust risk writes

$$\mathcal{R}_{\rho^2}(f) = \inf_{\lambda \geq 0} \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda)] . \quad (6.20)$$

These expressions for the risks are the bedrock of our analysis.

SKETCH OF PROOF. Our proof is built on two main parts: the first part is to obtain a concentration bound on the dual problems that crucially relies on a lower bound of the dual multiplier; the second part then consists in establishing such a lower bound. All the bounds are valid with high probability, and we drop the dependency on the confidence level δ of the theorems for simplicity.

For the first part of the proof (Appendix D.2), we assume that there is a deterministic lower-bound $\underline{\lambda} > 0$ on the optimal dual multiplier in (6.19) that holds with high-probability. As a consequence, we can restrict the range of λ in (6.19) to obtain:

$$\widehat{\mathcal{R}}_{\rho^2}(f) = \inf_{\lambda \geq \underline{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda)] \right\} \quad (6.21)$$

$$= \inf_{\lambda \geq \underline{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda)] - \lambda \frac{\mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda)] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda)]}{\lambda} \right\} \quad (6.22)$$

$$\geq \inf_{\lambda \geq \underline{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda)] - \lambda \sup_{\lambda' \geq \underline{\lambda}} \frac{\mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda')] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda')]}{\lambda'} \right\} \quad (6.23)$$

$$\geq \inf_{\lambda \geq \underline{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda)] - \lambda \bar{\rho}_n^2 \right\} \quad (6.24)$$

$$\geq \mathcal{R}_{\rho^2 - \bar{\rho}_n^2}(f) . \quad (6.25)$$

In the above, we used that the inner supremum, which is random, can be bounded by a deterministic and explicit quantity that we call $\bar{\rho}_n^2$, i.e.,

$$\bar{\rho}_n^2 \geq \sup_{\lambda' \geq \underline{\lambda}} \frac{\mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda')] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda')]}{\lambda'} \quad \text{with high probability.} \quad (6.26)$$

Hence, we obtain an upper-bound on the robust risk w.r.t. the true distribution with radius $\rho^2 - \bar{\rho}_n^2$. Moreover, we show that $\bar{\rho}_n^2 = \mathcal{O}(1/(\underline{\lambda}\sqrt{n}))$ which highlights the need for a precise lower bound $\underline{\lambda}$ to control the decrease in radius.

The second part of the proof thus consists in showing that the dual variable is indeed bounded away from 0, which means that the Wasserstein constraint is *sufficiently active*. We have to handle two cases differently:

- when ρ is small, i.e., close to ρ_n (Appendix D.3),
- when ρ is large, i.e., close to the critical radius ρ_c (Appendix D.4). Note that the additional Assumption 6.5 is required here: where we need to

control the behaviors of $f \in \mathcal{F}$ close to their maxima (see (6.18) for $\varepsilon = 0$ and small λ).

In both cases we obtain that $\underline{\lambda}$ scales as $1/\rho$ for the respective ranges of admissible radiuses. As a consequence $\bar{\rho}_n^2$ is bounded by $\rho\rho_n$ with $\rho_n = \mathcal{O}(1/\sqrt{n})$ and (6.25) becomes

$$\widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathcal{R}_{\rho(\rho-\rho_n)}(f), \quad (6.27)$$

which leads to our main results.

EXTENSION: UPPER AND LOWER BOUNDS ON THE EMPIRICAL ROBUST RISK. The proof that we sketched above actually shows that $\mathcal{R}_{\rho(\rho-\rho_n)}^\varepsilon(f)$ is a lower bound of $\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f)$. This proof technique also yields an upper bound by exchanging the roles of P and $\widehat{P}_n$.

Theorem 6.3. *In the setting of either Theorem 6.1, Theorem 6.2 or Theorem 6.6 (with $\varepsilon = 0$ or $\varepsilon > 0$), with probability at least $1 - \delta$, it holds that*

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_{\rho(\rho-\rho_n)}^\varepsilon(f) \leq \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \leq \mathcal{R}_{\rho(\rho+\rho_n)}^\varepsilon(f), \quad (6.28)$$

with $\rho_n = \mathcal{O}\left(\sqrt{\frac{1+\log 1/\delta}{n}}\right)$.

This result shows how two robust objectives w.r.t. P provide upper and lower bounds on the empirical robust risk, with only slight variations in the radius. Furthermore, when the number of data points n grows, both sides of the bound converge to the same quantity $\mathcal{R}_{\rho^2}^\varepsilon(f)$. Hence our generalization bounds of the form (6.27) are asymptotically tight.

The current proofs of this theorem and the previous ones rely on the cost being the squared Euclidean norm. This limitation was relaxed in the subsequent work (Le and Mallick, 2025), which builds on our approach and extends it in two directions. First, their results cover non-smooth loss functions, going beyond the smoothness required by Assumption 6.2. Second, they remove Assumptions 6.4 and 6.5 at the cost of non-explicit constants in the bounds. Together, these extensions establish similar generalization guarantees for WDRO under weaker conditions.

6.4 REGULARIZED WDRO

We now study WDRO with entropic regularization. We refer to Appendix C for the full paper Azizian et al. (2023a), which covers more general regularizers and unbounded sample spaces.

6.4.1 Regularizing the WDRO problem

The standard WDRO risks $\widehat{\mathcal{R}}_{\rho^2}(f)$ and $\mathcal{R}_{\rho^2}(f)$ involve a supremum over an infinite-dimensional ambiguity set, which makes them computationally challenging in general. A natural remedy, inspired by entropic regularization in optimal transport (OT) (Peyré et al., 2019, Chap. 4), is to penalize the supremum by a KL divergence with respect to a reference transport plan. This regularization yields a tractable dual objective (see § 6.4.2), and, as we show in § 6.4.4, enjoys similar generalization guarantees as the standard WDRO risks.

We now make this construction precise. For any transport plans $\pi, \bar{\pi} \in \mathcal{P}(\Xi \times \Xi)$, the KL divergence is defined by

$$\text{KL}(\pi|\bar{\pi}) = \begin{cases} \int \log \frac{d\pi}{d\bar{\pi}} d\pi & \text{when } \pi \ll \bar{\pi} \\ +\infty & \text{otherwise.} \end{cases} \quad (6.29)$$

Unlike in OT though, the choice of the reference measure in WDRO is not neutral and introduces a bias in the robust objective (Azizian et al., 2023a). For their theoretical convenience, we take reference measures that have Gaussian conditional distributions

$$\pi_\sigma(d\zeta|\xi) \propto \mathbb{1}_{\zeta \in \Xi} e^{-\frac{\|\xi - \zeta\|^2}{2\sigma^2}} d\zeta, \quad \text{for all } \xi \in \Xi, \quad (6.30)$$

where $\sigma > 0$ controls the spread of the second marginals, following Wang et al. (2021), Azizian et al. (2023a). Then, the regularized version of $\widehat{\mathcal{R}}_{\rho^2}(f)$ (WDRO empirical risk) is given by

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) := \sup_{\substack{\pi \in \mathcal{P}(\Xi \times \Xi), \pi_1 = \widehat{\mathbb{P}}_n \\ \mathbb{E}_{(\xi, \zeta) \sim \pi} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \leq \rho^2}} \mathbb{E}_{\xi \sim \pi_2} [f(\xi)] - \varepsilon \text{KL}(\pi | \pi_\sigma^n) \quad \text{with } \pi_\sigma^n = \widehat{\mathbb{P}}_n(d\xi) \pi_\sigma(d\zeta|\xi) \quad (6.31)$$

and similarly, the regularized version of $\mathcal{R}_{\rho^2}(f)$ is given by

$$\mathcal{R}_{\rho^2}^\varepsilon(f) := \sup_{\substack{\pi \in \mathcal{P}(\Xi \times \Xi), \pi_1 = \mathbb{P} \\ \mathbb{E}_{(\xi, \zeta) \sim \pi} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \leq \rho^2}} \mathbb{E}_{\xi \sim \pi_2} [f(\xi)] - \varepsilon \text{KL}(\pi | \pi_\sigma) \quad \text{with } \pi_\sigma = \mathbb{P}(d\xi) \pi_\sigma(d\zeta|\xi). \quad (6.32)$$

These regularized risks have been studied in terms of computational or approximation properties, but their statistical properties have not been investigated yet. The analysis we develop for WDRO estimators is general enough to carry over to these settings.

In (6.31) and (6.32), note finally that the regularization is added as a penalization in the supremum, rather than in the constraint. As in Wang et al. (2021), penalizing in the constraint leads to an ambiguity set defined by the regularized Wasserstein distance, that we introduce in § 6.4.4. We refer to Azizian et al. (2023a) for a unified presentation of the two penalizations.

We establish a dual formulation for this regularized problem in the next section, and study its approximation properties in § 6.4.3.

6.4.2 Dual formulation of the regularized WDRO problem

Duality plays a central role in both the theoretical analysis and computational schemes of WDRO. The following proposition extends the dual formulation of (6.19) and (6.20) to the regularized case.

Proposition 6.4. *Under Assumption 6.1, if the primal problem (6.31) is strictly feasible, then*

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\lambda \geq 0} \lambda \rho^2 + \varepsilon \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} \log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \exp \left(\frac{f(\zeta) - \lambda \|\xi - \zeta\|^2/2}{\varepsilon} \right) \right) \quad (6.33)$$

and similarly for $\mathcal{R}_{\rho^2}^\varepsilon(f)$ with P replacing $\widehat{P}_n$.

The interest of using entropy as a regularization appears when comparing this expression to the general dual (6.19): there is no inf over an auxiliary bivariate function φ , and the inner supremum is replaced by a smoothed log-sum-exp expression.

Proof. Follows from specializing [Theorem C.1](#) to the Gaussian reference measure (6.30) and the Euclidean squared cost, and setting the constraint regularization parameter to zero. ■

A key advantage of the entropic dual [Proposition 6.4](#) over the standard WDRO dual (6.19) is computational tractability. Standard WDRO requires specific model structure — typically linearity of the loss in the parameters — to admit a closed-form solution; by contrast, the log-sum-exp objective in [Proposition 6.4](#) is differentiable in the model parameters for any differentiable loss f , making it directly amenable to gradient-based optimization. These dual formulas are implemented in `skwdro` ([Vincent et al., 2026](#)), an open-source Python library for WDRO-based machine learning that we contributed to, which enables gradient-based training of standard and regularized WDRO estimators for arbitrary differentiable models.

Remark 8. [Blanchet and Kang \(2020\)](#) independently introduced a related dual smoothing technique, though in a restricted setting (finite Ξ with a uniform reference measure) and without the formal connection to primal KL regularization. Concurrently, [Wang et al. \(2021\)](#) proposed and analyzed entropic regularization of WDRO as an explicit primal formulation, with a focus on computational aspects; the present work complements theirs by providing a general theoretical treatment and approximation guarantees. After our work ([Azizian et al., 2023a](#)), the entropic regularization framework has subsequently been applied in a number of works, such as ([Piat et al., 2022](#), [Dapogny et al., 2023](#), [Song et al., 2023](#), [Wang and Xie, 2022](#), [Wang et al., 2022a](#), [Dapogny et al., 2025](#)).

6.4.3 Approximation of the regularized WDRO problem

We now study the behavior of the approximation error of the regularized problem (6.31) as the regularization parameter vanishes. We specify the cost $c(\xi, \zeta) = \frac{1}{2}\|\xi - \zeta\|^2$ (the Euclidean squared case used throughout) and the reference measure π_σ as in (6.30), with $\sigma > 0$ fixed such that $\mathbb{E}_{\pi_\sigma} c < \rho$.

Theorem 6.5 (Approximation for entropic regularization). *Let [Assumptions 6.1](#) and [6.2](#) hold, and suppose the objective f is Lipschitz continuous. Then, as $\varepsilon \rightarrow 0$ (with $\sigma > 0$ fixed),*

$$0 \leq \widehat{\mathcal{R}}_{\rho^2}(f) - \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \leq \mathcal{O}\left(d \varepsilon \log \frac{1}{\varepsilon}\right) \quad (6.34)$$

where $\bar{\lambda} := \frac{2 \sup_{\Xi} |f|}{\rho - \mathbb{E}_{\pi_\sigma} c}$ is an explicit dual bound.

This result is the analogue for WDRO of the quantitative bounds for entropic OT established in [Genevay et al. \(2019\)](#). The proof introduces a block approximation of the optimal transport plan following [Carlier et al. \(2017\)](#); the extension to WDRO requires an auxiliary regularized problem since the second marginal is not fixed. See [Theorem C.3](#) for the extended theorem with explicit constants.

6.4.4 Generalization guarantees for the regularized WDRO problem

The analysis that we develop for the standard WDRO estimators is general enough to also cover the regularized versions presented in § 6.4.1. We thus obtain the following [Theorem 6.6](#) which is the first generalization guarantee for regularized WDRO. This theorem is very similar to [Theorem 6.1](#) with still a couple of differences. First, the regularization leads to ambiguity sets defined in terms of $W_{2,\delta}(P, \cdot)$, the *regularized* Wasserstein distance to the true distribution P , defined, for some regularization parameter $\delta > 0$, as

$$W_{2,\delta}^2(P, Q) := \inf \left\{ \mathbb{E}_\pi \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] + \delta \text{KL}(\pi | \pi_\sigma) : \pi \in \mathcal{P}(\Xi \times \Xi), \pi_1 = P, \pi_2 = Q \right\}, \quad (6.35)$$

where π_σ appears in the definition of the regularized robust risk (6.32). Besides, the regularization allows us to avoid [Assumptions 6.4](#) and [6.5](#) to show our generalization result.

Theorem 6.6. *For $\sigma = \sigma_0 \rho$ with $\sigma_0 > 0$, $\varepsilon = \varepsilon_0 \rho$ with $\varepsilon_0 > 0$ such that $\varepsilon_0 / \sigma_0^2$ is small enough depending on $\mathcal{H}, P, \Xi$, there is an explicit constant ρ_c depending only on $\mathcal{H}, P$ and Ξ such that for all $\delta \in (0, 1)$ and $n \geq 1$, if*

$$\mathcal{O} \left(\sqrt{\frac{1 + \log 1/\delta}{n}} \right) \leq \rho \leq \frac{\rho_c}{2} - \mathcal{O} \left(\frac{1}{\sqrt{n}} \right), \quad \text{and} \quad \rho_c \geq \mathcal{O} \left(\frac{1}{n^{1/6}} + \left(\frac{1 + \log 1/\delta}{n} \right)^{1/4} \right), \quad (6.36)$$

then, there are $\delta = \mathcal{O}(\varepsilon \rho)$ and $\rho_n = \mathcal{O} \left(\sqrt{\frac{1 + \log 1/\delta}{n}} \right)$ such that, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{H}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \mathbb{E}_{\xi \sim Q} [f(\xi)] \quad \text{for all } Q \text{ such that } W_{2,\delta}^2(P, Q) \leq \rho(\rho - \rho_n). \quad (6.37)$$

Furthermore, when σ_0 and σ are small enough depending on P and Ξ , with probability $1 - \delta$,

$$\forall f \in \mathcal{H}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \mathbb{E}_{\xi \sim P} \mathbb{E}_{\zeta \sim \pi_\sigma(\cdot | \xi)} [f(\zeta)]. \quad (6.38)$$

The first part of the theorem, (6.37), guarantees that the empirical robust risk is an upper-bound on the loss even with some perturbations of the true distribution. As in OT, the regularization added to the Wasserstein metric induces a bias that may prevent $W_{2,\delta}^2(P, P)$ from being null. As a result, the second part of the theorem involves a smoothed version of the true risk: the empirical robust risk provides an *exact* upper-bound the true expectation of a convolution of the loss with π_σ .

A few additional comments are in order:

- Our result prescribes the scaling of the regularization parameters: ε and σ should be taken proportional to ρ .
- The critical radius ρ_c has a slightly more intricate definition, yet the same interpretation as in the standard WDRO case in [Remark 7](#); see [Appendix D.4.2](#).
- The regularized OT distances do not suffer from the curse of dimensionality ([Genevay et al., 2019](#)). However this property does not directly carry over to

regularized WDRO. Indeed, we cannot choose the same reference measure as in OT and we have to fix the measure π_σ , introducing a bias. As a consequence, we have to extend the analysis of the previous section to obtain the claimed guarantees that avoid the curse of dimensionality.

6.4.5 Examples

Our main theorems [Theorems 6.1](#), [6.2](#) and [6.6](#) involve a general class $\mathcal{H}$ of loss functions. We explain in this section how to instantiate our results in the important class of parametric models. We then illustrate this setting with both linear and non-linear examples and provide some numerical illustrations.

Let us consider the class of functions of the form

$$\mathcal{H} = \{\xi \mapsto g(\theta, \xi) : \theta \in \Theta\} \quad \text{with } g : \Theta \times \Xi \longrightarrow \mathbb{R} \quad (6.39)$$

where Θ , the parameter space, is a subset of $\mathbb{R}^p$ and Ξ , the sample space, is a subset of $\mathbb{R}^d$.

For instance, this covers the case of linear models of the form $g(\theta, \xi) = \ell(\langle \xi, \theta \rangle)$ with ℓ a convex loss. This class of models is studied by [Shafieezadeh-Abadeh et al. \(2019b\)](#), [Chen and Paschalidis \(2018\)](#) in a slightly different setting, where they obtain a closed form for the robust objective and then establish a generalization bound similar to [\(6.13\)](#).

Let us show how to instantiate our theorems in the case of [\(6.39\)](#).

- If g is twice continuously differentiable on a neighborhood of $\Theta \times \Xi$ with Θ and Ξ both compact, then [Assumption 6.2](#) is immediately satisfied. Therefore, [Theorem 6.6](#) can be readily applied and its generalization guarantee hold.
- As for [Assumption 6.4](#), it is equivalent to, for all $\theta \in \Theta$, $P(\nabla_\xi g(\theta, \xi) \neq 0) > 0$. Thus disregarding the degenerate case of $\nabla_\xi g(\theta, \xi)$ being null for P -almost every ξ (e.g., when the loss does not depend on ξ), we are in the setting of [Theorem 6.2](#).
- Satisfying [Assumption 6.5](#), needed for [Theorem 6.1](#), requires some problem-dependent developments, see the examples below. Note though that the second item of [Assumption 6.5](#) is implied by the parametric Morse-Bott property ([Arbel and Mairal, 2022](#)); see [Appendix D.1.5](#).

We discuss linear and non-linear examples of this framework. In light of the above, we focus our discussion on [Assumption 6.5](#). We first present the examples of linear models, [Examples 6.1](#) and [6.2](#), where the latter assumption is satisfied. We then consider several examples of nonlinear models: kernel regression ([Example 6.3](#)), smooth neural networks ([Example 6.4](#)) and families of invertible mappings ([Example 6.5](#)). In the appendix, we also provide numerical illustrations for linear models.

Example 6.1 (Logistic Regression). For a training sample $(x, y) \in \mathbb{R}^p \times \{-1, +1\}$, the logistic loss for a parameter $\theta \in \mathbb{R}^p$ is given by $\log(1 + e^{-y\langle x, \theta \rangle})$. It fits into our framework by defining $g(\theta, \xi) = \log(1 + e^{\langle \xi, \theta \rangle})$ with ξ playing the role of $-y \times x$. We assume that Θ is a compact set that does not include the origin, and, for the sake of simplicity, we take Ξ as a closed Euclidean ball, i.e., $\Xi = \overline{B}(0, r)$. We are going to show that [Assumption 6.5](#) is satisfied, and, for this, we need

the following elements. For any θ , the maximizer of $g(\theta, \cdot)$ over $\Xi = \overline{\mathbb{B}}(0, r)$ is reached at $\xi^* := \frac{r\theta}{\|\theta\|}$. Besides, for any $\xi \in \Xi$, it holds that

$$r^2 \geq \|\xi\|^2 = \|\xi^*\|^2 + 2\langle \xi^*, \xi - \xi^* \rangle + \|\xi - \xi^*\|^2, \quad (6.40)$$

so that, since $\|\xi^*\| = r$, we have

$$\langle \xi^*, \xi - \xi^* \rangle \geq \frac{1}{2} \|\xi - \xi^*\|^2. \quad (6.41)$$

We can now turn to the verification of [Assumption 6.5](#).

1. Take some $R > 0$ and some $\xi \in \Xi$ such that $\|\xi - \xi^*\| \geq R$. Then, [\(6.41\)](#) yields

$$\langle \theta, \xi \rangle - \langle \theta, \xi^* \rangle = \frac{\|\theta\|}{r} \langle \xi^*, \xi - \xi^* \rangle \leq -\frac{\|\theta\|}{2r} \|\xi - \xi^*\|^2 \leq -\frac{d(0, \Theta)R^2}{2r}. \quad (6.42)$$

Since $u \mapsto \log(1 + e^u)$ is increasing, this yields that $g(\theta, \xi) - g(\theta, \xi^*)$ is bounded away from 0 by a negative constant uniformly in θ . The first item of [Assumption 6.5](#) is thus satisfied.

2. Fix $\theta \in \Theta$; by Taylor expanding $u \mapsto \log(1 + e^u)$ around $\langle \theta, \xi^* \rangle$ we get

$$g(\theta, \xi) = g(\theta, \xi^*) + \frac{1}{1 + e^{-\langle \theta, \xi^* \rangle}} \langle \theta, \xi - \xi^* \rangle + \mathcal{O}(\langle \theta, \xi - \xi^* \rangle)^2, \quad (6.43)$$

where the big-O remainder is uniform over $\theta \in \Theta$. Using the first inequality in [\(6.42\)](#), we get for ξ close enough to ξ^*

$$g(\theta, \xi) \leq g(\theta, \xi^*) - \frac{1}{2(1 + e^{-\langle \theta, \xi^* \rangle})} \langle \theta, \xi - \xi^* \rangle \leq g(\theta, \xi^*) - \frac{\|\theta\|}{4r} \|\xi - \xi^*\|^2. \quad (6.44)$$

This shows that the second item of [Assumption 6.5](#) is satisfied locally around ξ^* . It can be made global by using the uniform Lipschitz-continuity of g , which introduces a term of the form $\frac{L}{6} \|\xi - \xi^*\|^3$.

Example 6.2 (Linear Regression). With samples of the form $\xi = (x, y) \in \mathbb{R}^p \times \mathbb{R}$ and parameters $\theta \in \mathbb{R}^p$, the loss is given by $g(\theta, \xi) = \frac{1}{2}(\langle \theta, x \rangle - y)^2$. Similarly to the previous example, we take Θ as a compact set of $\mathbb{R}^d$ that does not include the origin and Ξ of the form $\overline{\mathbb{B}}(0, r) \times [-r', r']$. The maximizers of $g(\theta, \cdot)$ on Ξ are $\xi_1^* = (r\theta/\|\theta\|, -r')$ and $\xi_2^* = (-r\theta/\|\theta\|, r')$. By symmetry, one can restrict to the case of ξ_1^* and $\langle \theta, x \rangle - y \geq 0$; the same rationale as above can then be applied.

Example 6.3 (Kernel Ridge Regression). Using a kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ with $\mathcal{X}$ compact and k smooth, for instance Gaussian or polynomial, we consider the following class of loss functions:

$$f(\theta, \xi) = \frac{1}{2} \left(\sum_{i=1}^m \alpha_i k(x, x_i) - y \right)^2 + \frac{\mu}{2} \|\alpha\|_2^2. \quad (6.45)$$

where $\xi = (x, y)$, Ξ is some compact subset of $\mathcal{X} \times \mathbb{R}$, $\theta = (\alpha_1, \dots, \alpha_m, x_1, \dots, x_m)$, $\Theta = A_m \times \mathcal{X}_m$, m is a fixed integer, A is a compact subset of $\mathbb{R}^m$, $\mathcal{X}_m$ can be any closed subset of $\mathcal{X}^m$ and $\mu \geq 0$ is the regularization parameter. A typical choice for $\mathcal{X}_m$ would be the datapoints of the training set. This class then fits into our

framework of parametric models above. Finally, further information on the kernel would be needed to ensure that [Assumption 6.5](#) is satisfied.

Example 6.4 (Smooth Neural Networks). Denote by $\mathcal{NN}(x, \theta, \sigma)$ a multi-linear perceptron that takes x as input, has weights θ and a smooth activation function σ , for instance the hyperbolic tangent or the Gaussian Error Linear Units (GELU). We choose $\ell(\hat{y}, y)$ a smooth loss function and we consider the loss $g(\theta, (x, y)) = \ell(\mathcal{NN}(x, \theta, \sigma), y)$ with $\theta \in \Theta$ some compact set. Provided that the inputs (x, y) lie in a compact set Ξ , this class fits the parametric framework above. Note that we require σ to be smooth, further work would be required for non-smooth activation functions.

Example 6.5 (Family of diffeomorphisms). Consider maps $h: \Xi \rightarrow \Xi$ and $(\theta, \xi) \in \Theta \times \Xi \mapsto h_\theta(\xi) \in \Xi$ and define the parametric loss $g(\theta, \xi) = h(h_\theta(\xi))$. Assume that these functions are twice differentiable, that h satisfies the second item of [Assumption 6.5](#) and that, for every $\theta \in \Theta$, h_θ has a inverse h_θ^{-1} which is also continuously differentiable in a neighborhood of $\Theta \times \Xi$.

As before, this setting fits into the framework above. We now show that [Assumption 6.5](#) is satisfied.

1. Since h is continuous, h satisfies the first item of [Assumption 6.5](#). It is satisfied by $\mathcal{H}$ as well thanks to h_θ^{-1} being Lipschitz-continuous in ξ uniformly in θ by compactness of $\Theta \times \Xi$.
2. Take C such that both h_θ and h_θ^{-1} are C -Lipschitz in ξ uniformly in θ . Since $\arg \max g(\theta, \cdot) = h_\theta^{-1}(\arg \max h)$, it holds that

$$\min_{\xi^* \in \arg \max h} \|h_\theta(\xi) - \xi^*\| = \min_{\xi^* \in \arg \max g(\theta, \cdot)} \|h_\theta(\xi) - h_\theta(\xi^*)\| \quad (6.46)$$

which lies between $C^{-1} \min_{\xi^* \in \arg \max g(\theta, \cdot)} \|\xi - \xi^*\|$ and $C \min_{\xi^* \in \arg \max g(\theta, \cdot)} \|\xi - \xi^*\|$. Combined with h satisfying the second item of [Assumption 6.5](#), this shows that g satisfies this condition as well.

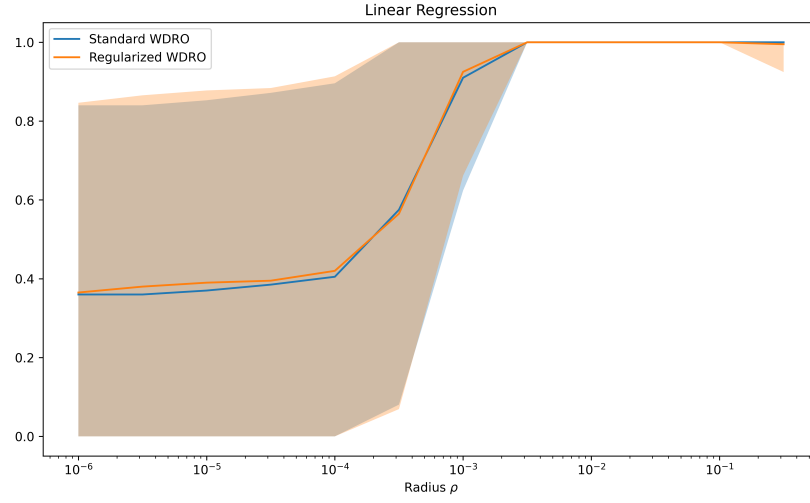
We conclude this section with numerical illustrations on the logistic and linear regression models of [Examples 6.1](#) and [6.2](#). Following [Esfahani and Kuhn \(2018, §7.2.A\)](#), we estimate the probability

$$\mathbb{P}^{\otimes n} \left(\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(g(\widehat{\theta}_n, \cdot)) \geq \mathbb{E}_{\xi \sim P} \left[g(\widehat{\theta}_n, \xi) \right] \right) \quad \text{where} \quad \widehat{\theta}_n = \arg \min_{\theta \in \Theta} \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(g(\theta, \cdot)), \quad (6.47)$$

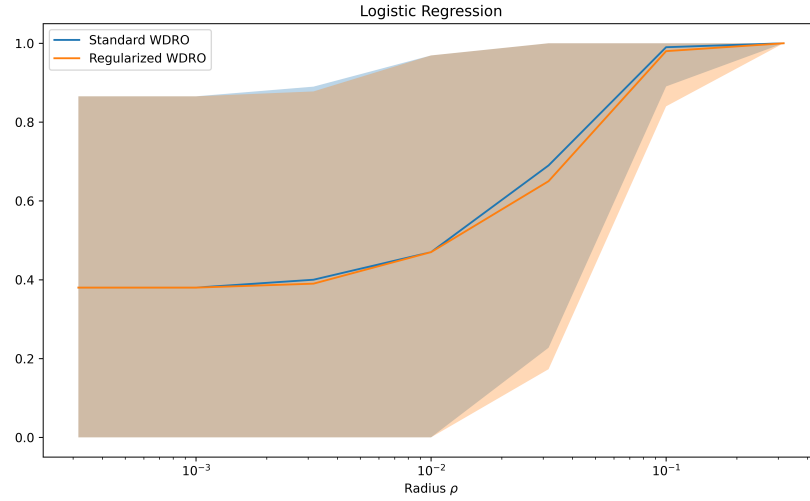
where $\mathbb{P}^{\otimes n}$ denotes the distribution of the training set $(\xi_i)_{1 \leq i \leq n}$ with $\xi_i \sim P$ i.i.d.. As predicted by [Theorems 6.1](#) and [6.6](#), this probability is close to 1 for ρ large enough, for both models and both standard and regularized WDRO.

6.5 CONCLUSION

In this work, we provide generalization guarantees for WDRO models that improve over existing literature in the following aspects: our results avoid the curse of dimensionality, provide exact upper bounds without spurious error terms, and allow for distribution shifts during testing. We obtained these bounds through the development of an original concentration result on the dual of WDRO. Our framework is general enough to cover regularized versions of the WDRO problem: they enjoy similar generalization guarantees as standard WDRO, with less restrictive assumptions.



Estimates of the probability $P^{\otimes n} \left(\widehat{\mathcal{R}}_{\rho^2}^{\epsilon}(f(\hat{\theta}_n, \cdot)) \geq \mathbb{E}_P[f(\hat{\theta}_n, \xi)] \right)$ where $\hat{\theta}_n$ is the robust model with radius ρ for the linear regression model (Example 3.7). θ has dimension $d = 10$, $n = 1000$ synthetic training samples are used, σ and ϵ are chosen proportional to ρ following Theorem 6.6. For each value of ρ , we sample 200 training datasets and solve the WDRO problem on each of them, to obtain an estimate of the probability above. The solid line is the average over these 200 results, while the shaded area represents the standard deviation. As predicted by Theorems 6.1 and 6.6, we observe that for ρ large enough, the probability that the robust loss on the training set upper bounds the true risk is almost 1. We also observe that standard and regularized WDRO have almost identical generalization behaviours. The WDRO problems are solved by LBFGS-B combined with the formulas of Example 2 of Wang et al. (2021).



Estimates of the probability $P^{\otimes n} \left(\widehat{\mathcal{R}}_{\rho^2}^{\epsilon}(f(\hat{\theta}_n, \cdot)) \geq \mathbb{E}_P[f(\hat{\theta}_n, \xi)] \right)$ where $\hat{\theta}_n$ is the robust model with radius ρ for the logistic regression model (Example 3.6). θ has dimension $d = 5$, $n = 500$ synthetic training samples are used, σ and ϵ are chosen proportional to ρ following Theorem 6.6. For each value of ρ , we sample 100 training datasets and solve the WDRO problem on each of them, to obtain an estimate of the probability above. The solid line is the average over these 100 results, while the shaded area represents the standard deviation. As for the previous plot and as predicted by Theorems 6.1 and 6.6, we observe that for ρ large enough, the probability that the robust loss on the training set upper bounds the true risk is almost 1. We also observe that standard and regularized WDRO have almost identical generalization behaviours. The standard WDRO problem is solved using the algorithm of Blanchet et al. (2022b) while the regularized problem is solved using LBFGS-B combined with an explicit expression of the inner integral in the robust loss.

Though this work already yields some practical insights on how ρ should be chosen, the crucial role played by the radius of the Wasserstein ball for generalization calls for a principled and efficient procedure to select it.

THE RATE OF CONVERGENCE OF BREGMAN PROXIMAL METHODS: LOCAL GEOMETRY VS. REGULARITY VS. SHARPNESS

7.1	Introduction	113
7.2	Problem setup and preliminaries	115
7.2.1	Blanket assumptions	115
7.2.2	Bregman proximal methods	116
7.3	Motivating examples	118
7.4	The Legendre exponent and convergence rate analysis	120
7.4.1	The Legendre exponent	121
7.4.2	Convergence rate analysis	122
7.4.3	Proof of Theorem 7.2	124
7.5	Finer results for linearly constrained problems	129
7.5.1	Motivating examples, redux	129
7.5.2	Linearly constrained problems	130
7.5.3	Convergence rate analysis	131
7.5.4	Proof of Theorem 7.9	134
7.6	Mirror descent on stochastic variational inequalities	138
7.6.1	The Stochastic optimistic mirror descent algorithm	138
7.6.2	Analysis and results	139
7.6.3	Main ideas of the proof	141
7.7	Conclusion	144

This chapter is a combination of two articles: [Azizian et al. \(2024d\)](#) and [Azizian et al. \(2021b\)](#). This line of works proposes a refined analysis of Bregman proximal methods through the so-called Legendre exponent. [Azizian et al. \(2024d\)](#) introduces this notion and focuses on the deterministic setting while [Azizian et al. \(2021b\)](#), presented in § 7.6.1, extends the analysis to the stochastic setting.

7.1 INTRODUCTION

Bregman proximal methods (BPMs) have a long and rich history in optimization, going back at least to the introduction of mirror descent by Nemirovski & Yudin ([Nemirovski and Yudin, 1983](#)). In plain terms, BPMs are first-order (constrained) optimization algorithms that forego Euclidean projections in favor of a more sophisticated “prox-mapping” that minimizes a certain distance-like functional known as the Bregman divergence ([Nemirovski and Yudin, 1983](#), [Chen and Teboulle, 1993](#), [Bregman, 1967](#), [Kiwiel, 1997](#)). When this Bregman divergence is the Euclidean distance squared, one recovers the standard projection-based methods; other than that, depending on the problem’s feasible region, different Bregman setups lead to a diverse collection of algorithms, from exponentiated

gradient descent on the simplex (Nemirovski and Yudin, 1983, Beck and Teboulle, 2003, Auer et al., 2002), to matrix multiplicative weights on the positive-semidefinite cone Tsuda et al. (2005), Kakade et al. (2012), variants of Karmarkar’s affine scaling algorithm for linear programs Vanderbei et al. (1986), etc.

One of the most appealing features of BPMs is that they achieve almost dimension-free convergence rates in problems with a convex structure and a favorable geometry – such as the L^1 ball, the spectraplex, second-order cones, etc. Bubeck (2015), Nesterov (2009), Beck and Teboulle (2003). This is owed to a delicate interplay between the algorithms’ non-Euclidean update scheme and the global geometry of the problem’s domain. However, these (almost) dimension-free guarantees also come with some strings attached: they do not concern the sequence of iterates generated by the method, but only its time average (or, through the same, “regret-based” analysis, the method’s “best iterate”); in this way, the best guarantee that can be achieved after t iterations is $\mathcal{O}(1/t)$. In terms of oracle complexity, this is sufficient for problems that are not strongly convex / strongly monotone, but if one targets finer, geometric convergence rates, the inherent averaging involved in regret-based guarantees is hard to compensate. And, on the other extreme, if the problem is not convex / monotone to begin with, iterate averaging does not provide any quantifiable benefits whatsoever, so it becomes crucial to study the actual trajectory of the method.

OUR CONTRIBUTIONS In this general setting, our work quantifies the last-iterate convergence rate of Bregman proximal methods as a function of the Bregman divergence defining the method and the local geometry that it induces. To treat this question in as general a manner as possible, we focus on VI problems of the form

$$\text{Find } x^* \in \mathcal{X} \text{ such that } \langle g(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{X}, \quad (\text{VI})$$

where $\mathcal{X}$ is a closed convex subset of a finite-dimensional normed space $\mathcal{V}$, and $g: \mathcal{X} \rightarrow \mathcal{V}^*$ is a (possibly non-monotone) single-valued operator on $\mathcal{X}$ with values in $\mathcal{V}^*$, the dual of $\mathcal{V}$. This problem is a staple of many areas of mathematical programming, game theory and data science, as it provides a template for “optimization beyond minimization” – i.e., for problems where finding an optimal solution does not necessarily involve minimizing a loss function. In particular, in addition to standard minimization problems – which are recovered when $g = \nabla f$ for some smooth function f – the general formulation (VI) includes saddle-point problems, games, complementarity problems, etc.; for an introduction, see Facchinei and Pang (2003) and references therein.

In this broad context, we examine the rate of convergence of a wide class of Bregman proximal methods to local solutions of (VI) that satisfy a second-order sufficient condition. Specifically, the class of algorithms we consider includes as special cases (i) the original mirror descent (MD) algorithm of Nemirovski and Yudin (1983); (ii) the mirror-prox (MP) method of Nemirovski (2004) – which has the same update structure as the Bregman-based algorithm of Auslender and Teboulle (2005) and contains as a special case the extra-gradient (EG) algorithm of Korpelevich (1976); (iii) the so-called optimistic mirror descent (OMD) method of Rakhlin and Sridharan (2013) – itself a Bregman analogue of the modified Arrow-Hurwicz algorithm of Popov (1980); etc.

Our first finding is a crisp characterization of last-iterate convergence rate of BPMs in terms of the local geometry induced by the underlying Bregman function near a given solution of (VI). We make this dependence precise via the notion of the *Legendre exponent*, a regularity measure for Bregman methods that we introduce, which can roughly be described as the logarithmic ratio of the volume of a Euclidean ball to that of a Bregman ball of the same radius. For example, Euclidean methods have a Legendre exponent of $\beta = 0$ and they converge at a linear rate; entropic methods have a Legendre exponent of $\beta = 1/2$ at boundary points, and they converge at a rate of $\mathcal{O}(t^{-1})$; more generally, as we show in [Theorem 7.2](#), methods with a Legendre exponent $\beta > 0$ converge at a rate of $\mathcal{O}(t^{1-1/\beta})$. The Euclidean regime ($\beta = 0$) is perfectly aligned with existing results for the geometric last-iterate convergence rate of the EG algorithm and its variants ([Gidel et al., 2019](#), [Malitsky, 2015](#), [Hsieh et al., 2019](#), [Mokhtari et al., 2020](#)). By contrast, the Legendre regime ($\beta > 0$) indicates a significant drop in the algorithm's last-iterate convergence speed, even though ergodic convergence rates [Nesterov \(2004\)](#) and results for bilinear games [Wei et al. \(2021\)](#) might suggest otherwise.

Subsequently, motivated by applications to game theory and linear programming, we take a closer look at the convergence rate of BPMs across the constraints that are active at a solution x^* of (VI) depending on the position of $g(x^*)$ relative to said constraints. This analysis reveals that Bregman proximal methods have a particularly fine structure: along *sharp directions* (i.e., constraints along which $g(x^*)$ is strictly inward-pointing), BPMs converge (i) at a rate of $\mathcal{O}(1/t^{1/(2\beta-1)})$ if $1/2 < \beta < 1$; (ii) at a *geometric rate* if $0 < \beta \leq 1/2$ (e.g., for entropic methods); and (iii) in a *finite* number of iterations if $\beta = 0$ (cf. [Theorem 7.9](#)). Thus, even though the estimates of [Theorem 7.2](#) are, in general tight, the actual convergence rate of a Bregman method along different coordinates / constraints could be starkly different – and, in fact, dramatically faster if the solution under study is itself sharp.

7.2 PROBLEM SETUP AND PRELIMINARIES

In the rest of our paper $\mathcal{V}$ will denote an n -dimensional real space with norm $\|\cdot\|$ and $\mathcal{X}$ will be a closed convex subset thereof. We will also write $\mathcal{Y} := \mathcal{V}^*$ for the dual of $\mathcal{V}$, $\langle y, x \rangle$ for the canonical pairing between $y \in \mathcal{Y}$ and $x \in \mathcal{V}$, and $\|y\|_* := \max\{\langle y, x \rangle : \|x\| \leq 1\}$ for the induced dual norm on $\mathcal{Y}$.

7.2.1 Blanket assumptions

Throughout the sequel, we will make the following assumptions for the defining vector field $g: \mathcal{X} \rightarrow \mathcal{Y}$ of (VI) and the solution $x^* \in \mathcal{X}$ under study:

Assumption 7.1 (Lipschitz continuity). The vector field g is L -Lipschitz continuous, i.e.,

$$\|g(x') - g(x)\|_* \leq L\|x' - x\| \quad \text{for all } x, x' \in \mathcal{X}. \quad (\text{LC})$$

Assumption 7.2 (Second-order sufficiency). There exists a convex neighborhood $\mathcal{B}$ of x^* in $\mathcal{X}$ and a positive constant $\mu > 0$ such that

$$\langle g(x) - g(x^*), x - x^* \rangle \geq \mu\|x - x^*\|^2 \quad \text{for all } x \in \mathcal{B}. \quad (\text{SOS})$$

In general, [Assumption 7.2](#) guarantees that x^* is the unique solution of (VI) in $\mathcal{B}$; we illustrate this in two special cases of interest:

1. *Minimization problems:* suppose that $g = \nabla f$ for some Lipschitz smooth objective function f on $\mathcal{X}$. Then, [Assumption 7.2](#) implies that f grows (at least) quadratically along every ray emanating from x^* , i.e., $f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^* \rangle + (\mu/2)\|x - x^*\|^2 = \Omega(\|x - x^*\|^2)$ for all $x \in \mathcal{B}$, implying in particular that x^* is an isolated minimizer of f .
2. *Min-max problems:* suppose that $\mathcal{X}$ factorizes as $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ for suitable factor sets $\mathcal{X}_1$, and $\mathcal{X}_2$, let $\mathcal{L}: \mathcal{X} \rightarrow \mathbb{R}$ be a smooth function on $\mathcal{X}$, and write $g = (\nabla_{x_1} \mathcal{L}, -\nabla_{x_2} \mathcal{L})$ for the min-max gradient of $\mathcal{L}$ (with respect to $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ respectively). Then, any solution $x^* = (x_1^*, x_2^*)$ of (VI) that satisfies [Assumption 7.2](#) enjoys the local growth bounds $\mathcal{L}(x_1, x_2^*) - \mathcal{L}(x_1^*, x_2^*) = \Omega(\|x_1 - x_1^*\|^2)$ and $\mathcal{L}(x_1^*, x_2) - \mathcal{L}(x_1^*, x_2^*) = \Omega(\|x_2 - x_2^*\|^2)$, implying in turn that x^* is an isolated, hyperbolic saddle-point of $\mathcal{L}$.

More examples satisfying (SOS) include strict Nash equilibria in finite games [Fudenberg and Tirole \(1991\)](#), deterministic Nash policies in (generic) stochastic games [Solan and Vieille \(2015\)](#), etc. Overall, [Assumptions 7.1](#) and [7.2](#) apply to a very wide range of problems, so we will treat them as blanket assumptions throughout.

7.2.2 Bregman proximal methods

As we discussed in the introduction, the main algorithmic template that we will examine for solving (VI) is a general class of first-order algorithms known as *Bregman proximal methods* (BPMs). The defining ingredient of this class is the notion of *Bregman regularizer*, which we define below as follows:

Definition 7.3 (Bregman regularizers and related notions). A proper, lower semi-continuous, strictly convex function $h: \mathcal{V} \rightarrow \mathbb{R} \cup \{\infty\}$ is a *Bregman regularizer* on $\mathcal{X}$ if

1. h is supported on $\mathcal{X}$, i.e., $\text{dom } h = \mathcal{X}$.
2. The subdifferential of h admits a *continuous selection*, i.e., there exists a continuous mapping $\nabla h: \text{dom } \partial h \rightarrow \mathcal{Y}$ such that $\nabla h(x) \in \partial h(x)$ for all $x \in \text{dom } \partial h$.
3. h is *locally strongly convex* relative to $\|\cdot\|$, i.e., for any compact set $\mathcal{K} \subseteq \text{dom } h$, we have

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \frac{\kappa}{2} \|x' - x\|^2 \quad (7.1)$$

for some $\kappa > 0$ and for all $x \in \mathcal{K} \cap \text{dom } \partial h$, $x' \in \mathcal{K}$.

The set $\mathcal{X}_h := \text{dom } \partial h$ will be referred to as the *prox-domain* of h . In addition, we also define the *Bregman divergence* of h as

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x \rangle \quad \text{for all } x \in \mathcal{X}_h, p \in \mathcal{X} \quad (7.2)$$

and the induced *prox-mapping* as

$$P_x(y) = \arg \min_{x' \in \mathcal{X}} \{ \langle y, x - x' \rangle + D(x', x) \} \quad \text{for all } x \in \mathcal{X}_h, y \in \mathcal{Y}. \quad (7.3)$$

Remark 9. Examples of Bregman regularizers are given in [§ 7.3](#), where we also take an in-depth look at their properties. For our analysis and results, we will assume for convenience that h is 1-strongly convex in a suitable neighborhood $\mathcal{D}$ of x^* which will be understood from the context; this can always be achieved by rescaling h , so there is no loss of generality. To avoid technicalities, we will

also tacitly assume that $P_x(y)$ is well-defined whenever it is called (this is always the case if, for example, h is coercive or ∇h is invertible).

Given a Bregman regularizer on $\mathcal{X}$, the general class of *Bregman proximal methods* (BPMs) that we will consider is defined via the generic recursion

$$x_{t+1/2} = P_{x_t}(-\gamma_t g_t) \quad x_{t+1} = P_{x_t}(-\gamma_t g_{t+1/2}) \quad (\text{BPM})$$

where (i) $t = 1, 2, \dots$ denotes the method's iteration counter; (ii) $\gamma_t > 0$ is a (non-increasing) step-size sequence; (iii) g_t and $g_{t+1/2}$ are sequences of "oracle signals" that we discuss in detail below. In terms of vocabulary, the iterates $x_t, t = 1, 2, \dots$, will be referred to as the "*base states*" of the method, while the "half-iterates" $x_{t+1/2}, t = 1, 2, \dots$, will be referred to as the method's "*leading states*". Finally, in terms of initialization, we will take for convenience $x_1 = x_{1/2}$.

Now, regarding the sequence of oracle signals g_t and $g_{t+1/2}$ defining (BPM), we will assume throughout that

$$g_{t+1/2} = g(x_{t+1/2}) \quad \text{for all } t = 1, 2, \dots \quad (7.4)$$

i.e., (BPM) generates a new base state x_{t+1} by taking a Bregman proximal step from x_t with oracle input from the leading state $x_{t+1/2}$. By contrast, the leading state itself can be generated in a number of different ways from x_t , depending on the definition of g_t :

Assumption 7.4. For all $t = 1, 2, \dots$, the oracle signal g_t is of the form:

$$g_t = a g(x_t) + b g(x_{t-1/2}) \quad (7.5)$$

for some $a, b \in [0, 1]$ with $a + b \leq 1$ and $a + b = 1$ if $b > 0$.¹

For concreteness, we illustrate below three archetypal Bregman methods that serve as the backbone of the above framework:

1. *Mirror descent*: following Nemirovski and Yudin (1983), Beck and Teboulle (2003), Nemirovski et al. (2009), the *mirror descent* algorithm proceeds recursively as $x^+ = P_x(-\gamma g(x))$, so it can be recovered from (BPM) by taking

$$a = 0, b = 0 \quad \text{or, equivalently} \quad g_t = 0 \quad \text{for all } t = 1, 2, \dots \quad (\text{MD})$$

2. *Mirror-prox*: following Nemirovski (2004), Juditsky et al. (2011), the *mirror-prox* algorithm corresponds to the choice

$$a = 1, b = 0 \quad \text{or, equivalently} \quad g_t = g(x_t) \quad \text{for all } t = 1, 2, \dots \quad (\text{MP})$$

3. *Optimistic mirror descent*: originally due to Popov (1980) (in the Euclidean case) and Chiang et al. (2012), Rakhlin and Sridharan (2013) (for the general case), the *optimistic mirror descent* algorithm is obtained by setting

$$a = 0, b = 1 \quad \text{or, equivalently} \quad g_t = g(x_{t-1/2}) \quad \text{for all } t = 1, 2, \dots \quad (\text{OMD})$$

¹ Note that the requirement " $a + b = 1$ if $b > 0$ " is only intended to ease notation and does not lead to a loss in generality: if $b > 0$, we can always rescale γ_t by $a + b$ so the condition $a + b = 1$ is satisfied automatically.

These three algorithms are the most widely studied Bregman methods in the literature, so we will use them as running examples throughout.

7.3 MOTIVATING EXAMPLES

We now proceed to take a closer look at some commonly used Bregman regularizers (and the induced prox-mappings) with the goal of determining the rate of convergence of the associated Bregman method. For concreteness, we focus on one-dimensional problems where $\mathcal{X}$ is the closed interval $[0, \infty)$ or $[-1, 1]$ and g is the affine vector field

$$g(x) = x - x^*, \quad x \in \mathbb{R}, \quad (7.6)$$

for different choices of $x^* \in \mathbb{R}$ (typically a boundary point of $\mathcal{X}$). To streamline our presentation, we will only examine the mirror descent recursion (MD) with constant step-size schedules $\gamma_t \equiv \gamma$ for some $\gamma > 0$. In this case, we obtain the scheme

$$x_{t+1} = F(x_t) \quad \text{with} \quad F(x) = P_x(-\gamma g(x)), \quad (7.7)$$

and we will examine the convergence speed of x_t by analyzing the behavior of F near x^* . Our focus will be on cases where x^* is a boundary point of $\mathcal{X}$. This will demonstrate the variety of behaviours Bregman divergence can exhibit.

Example 7.1 (Euclidean regularization). We begin with the quadratic regularizer $h(x) = x^2/2$ for $x \in \mathcal{X} = [0, \infty)$. In this case, noting that $h'(x) = x$, we have:

$$\begin{aligned} a) \text{ Prox-domain:} & \quad \mathcal{X}_h = \mathcal{X} \\ b) \text{ Bregman divergence:} & \quad D(p, x) = (p - x)^2/2 \\ c) \text{ Prox-mapping:} & \quad P_x(y) = [x + y]_+ \end{aligned} \quad (7.8)$$

Consider now the case $x^* = 0$, i.e., $g(x) = x$. Then, for $\gamma \in (0, 1)$, the update (7.7) becomes

$$F(x) = x - \gamma x = (1 - \gamma)x \quad \text{for all } x \geq 0 \quad (7.9)$$

i.e., F is contracting. We thus conclude that x_t converges to $x^* = 0$ at a geometric rate, viz.

$$D(x^*, x_t) = \frac{1}{2}x_t^2 = \Theta((1 - \gamma)^{2t}) \quad \text{or, in absolute value, } |x_t - x^*| = \Theta((1 - \gamma)^t). \quad \heartsuit$$

Example 7.2 (Entropic regularization). Another popular choice when $\mathcal{X} = [0, \infty)$ is the entropic regularizer $h(x) = x \log x$ Beck and Teboulle (2003), Shalev-Shwartz (2011), Bauschke et al. (2017). In this case, $h'(x) = 1 + \log x$, which gives

$$\begin{aligned} a) \text{ Prox-domain:} & \quad \mathcal{X}_h = \text{ri } \mathcal{X} = (0, \infty) \\ b) \text{ Bregman divergence:} & \quad D(p, x) = p \log(p/x) + x - p \\ c) \text{ Prox-mapping:} & \quad P_x(y) = x \exp(y). \end{aligned} \quad (7.10)$$

Now, taking $g(x) = x$ as in the previous example, the update rule (7.7) becomes

$$F(x) = x \exp(-\gamma x) = x(1 - \gamma x + o(x)) = x - \gamma x^2 + o(x^2) \quad \text{as } x \rightarrow 0. \quad (7.11)$$

In contrast to (7.9), we now have $F(x) \sim x$ instead of $(1 - \gamma)x$, so F is no longer a contraction. Instead, the iterates of (7.11) may be analyzed by means of the following lemma:

Lemma 7.1. *Suppose that $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ admits the asymptotic expansion*

$$f(x) = x - \lambda x^{1+r} + o(x^{1+r}) \quad \text{as } x \rightarrow 0 \quad (7.12)$$

for positive constants $\lambda, r > 0$. Then, for $u_1 > 0$ small enough, the sequence $u_{t+1} = f(u_t)$, $t = 1, 2, \dots$, converges to 0 at a rate of $u_t \sim (\lambda r t)^{-1/r}$.

Thanks to this lemma (which we prove in Appendix E), we readily conclude that x_t converges to 0 at a rate of $D(x^*, x_t) = x_t = |x_t - x^*| \sim 1/(\gamma t)$. $\heartsuit$

Example 7.3 (Fractional power). Take $\mathcal{X} = [0, \infty)$ and $g(x) = x$ as in Examples 7.1 and 7.2 above. Then, for a given $q > 0$, $q \neq 1$, the fractional power regularizer – or Tsallis entropy – on $\mathcal{X}$ is defined as $h(x) = [q(1 - q)]^{-1}(x - x^q)$ (Tsallis, 1988, Alvarez et al., 2004, Mertikopoulos and Sandholm, 2016). For this choice of regularizer, we have $h'(x) = (1 - qx^{q-1})/[q(1 - q)]$, and a series of direct calculations gives:²

$$\begin{aligned} a) \text{ Prox-domain:} & \quad \mathcal{X}_h = (0, \infty) \text{ if } q \in (0, 1) \text{ and } \mathcal{X}_h = [0, \infty) \text{ if } q > 1 \\ b) \text{ Bregman divergence:} & \quad D(p, x) = \frac{x^q - p^q}{q(1 - q)} - x^{q-1} \frac{x - p}{1 - q} \\ c) \text{ Prox-mapping:} & \quad P_x(y) = [x^{q-1} - (1 - q)y]^{\frac{1}{q-1}} \quad \text{for } q \in (0, 1). \end{aligned} \quad (7.13)$$

Now, when applied to $g(x) = x$, the fractional power variant of (7.7) for $q \in (0, 1)$ gives

$$F(x) = x [1 + \gamma(1 - q)x^{2-q}]^{1/(q-1)} = x - \gamma x^{3-q} + o(x^{3-q}) \quad \text{as } x \rightarrow 0. \quad (7.14)$$

Hence, by Lemma 7.1, we conclude that x_t converges to 0 at a rate of

$$D(x^*, x_t) = \Theta(t^{-q/(2-q)}) \quad \text{or, in absolute value, } |x_t - x^*| = \Theta(t^{-1/(2-q)}). \quad \heartsuit$$

Example 7.4 (Hellinger distance). Our last example concerns the Hellinger regularizer $h(x) = -\sqrt{1 - x^2}$ on $\mathcal{X} = [-1, 1]$. Since $h'(x) = x/\sqrt{1 - x^2}$, we readily obtain the following:

$$\begin{aligned} a) \text{ Prox-domain:} & \quad \mathcal{X}_h = \text{ri } \mathcal{X} = (-1, 1) \\ b) \text{ Bregman divergence:} & \quad D(p, x) = \frac{1 - px - \sqrt{(1 - p^2)(1 - x^2)}}{\sqrt{1 - x^2}} \\ c) \text{ Prox-mapping:} & \quad P_x(y) = \frac{x + y\sqrt{1 - x^2}}{\sqrt{1 - x^2 + (x + y\sqrt{1 - x^2})^2}}. \end{aligned} \quad (7.15)$$

In this case, taking $g(x) = x$ as per the previous examples, yields

$$F(x) = \frac{x - \gamma x \sqrt{1 - x^2}}{\sqrt{1 - x^2 + (x - \gamma x \sqrt{1 - x^2})^2}} \sim x - \gamma x \quad \text{as } x \rightarrow 0, \quad (7.16)$$

² Strictly speaking, the expression we provide for $P_x(y)$ is only valid when $y < x^{q-1}/(1 - q)$. The reason for this is that the prox-mapping $P_x(y)$ is not well-defined for all values of y . This detail is not important in the calculations that follow, so we disregard it for now.

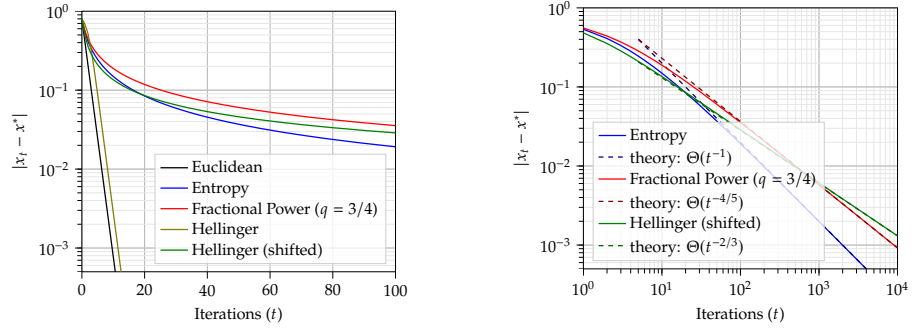


Figure 22: The rate of convergence of (MD) in Examples 7.1–7.4. The Euclidean and shifted Hellinger regularizers lead to a geometric rate (see left figure); all other examples converge at a polynomial rate.

i.e., x_t converges to $x^* = 0$ at a geometric rate, as in Example 7.1. On the other hand, if we consider the shifted operator $g(x) = x + 1$, a somewhat tedious calculation (which we detail Appendix E.5) gives the following Taylor expansion near $x^* = -1$:

$$F(x) = x^* + (x - x^*) - 2\sqrt{2}\gamma(x - x^*)^{5/2} + o\left((x - x^*)^{5/2}\right). \quad (7.17)$$

Hence, by Lemma 7.1, we conclude that x_t converges to $x^* = -1$ at a rate of

$$D(x^*, x_t) = \Theta(t^{-1/3}) \text{ or, in absolute value, } |x_t - x^*| = \Theta(t^{-2/3}). \quad \clubsuit$$

Albeit one-dimensional, the above examples provide a representative view of the geometry of Bregman proximal methods near a solution. Specifically, they show that the divergence induced by a given regularizer may exhibit a very different behavior at the boundary of $\mathcal{X}$: when x^* is a boundary point, $D(x^*, x)$ grows as $\Theta(\|x - x^*\|^2)$ in the Euclidean case, as $\Theta(\|x - x^*\|)$ for the negative entropy, and, more generally, as $\Theta(\|x - x^*\|^q)$ for the q -th power regularizer. As a result, when used as a measure of convergence, it is important to rescale the Bregman divergence in order to avoid inflating – or *deflating* – an algorithm’s rate of convergence.

Nonetheless, even if we take this rescaling into account, different instances of (MD) may lead to completely different rates of convergence. Specifically, in terms of absolute values, we observe a geometric rate in the Euclidean and shifted Hellinger cases, a rate of $\Theta(1/t)$ for the negative entropy, and a rate of $\Theta(1/t^{1/(2-q)})$ for the q -th power regularizer (cf. Figure 22 above). This is due to the different first-order behavior of the iterative update map $x \leftarrow F(x)$ that underlies (MD), which is itself intimately related to the growth rate of the Bregman divergence near a solution x^* of (VI). We make this relation precise in the next section.

7.4 THE LEGENDRE EXPONENT AND CONVERGENCE RATE ANALYSIS

Our goal in this section is to provide a precise link between the geometry induced by a Bregman regularizer near a solution and the convergence rate of the associated Bregman proximal method. The key notion in this regard is that of the *Legendre exponent*, which we define and discuss in detail below.

7.4.1 The Legendre exponent

Our starting point is the observation that, without loss of generality, the local strong convexity requirement for h can be expressed as

$$D(p, x) \geq \frac{1}{2} \|p - x\|^2 \quad \text{for all } x \in \mathcal{X}_h \text{ sufficiently close to } p. \quad (7.18)$$

Qualitatively, this means that the convergence topology induced by the Bregman divergence of h on $\mathcal{X}$ is *at least as fine* as the ambient norm topology: if a sequence $x_t \in \mathcal{X}_h$, $t = 1, 2, \dots$, converges to $p \in \mathcal{X}$ in the Bregman sense ($D(p, x_t) \rightarrow 0$), it also converges in the ambient norm topology ($\|x_t - p\| \rightarrow 0$). On the other hand, from a quantitative standpoint, the rate of this convergence could be quite different: as we saw in the previous section, the reverse inequality $D(p, x) = \mathcal{O}(\|p - x\|^2)$ may fail to hold, in which case $\sqrt{D(p, x_t)}$ and $\|x - x_t\|$ would exhibit a different asymptotic behavior.

To quantify this gap, we use the notion of the *Legendre exponent*, that we introduce in these series of works.

Definition 7.5. Let h be a Bregman regularizer on $\mathcal{X}$. The *Legendre exponent* of h at $p \in \mathcal{X}$ is defined as

$$\beta_h(p) := \inf \left\{ \beta \in [0, 1] : \limsup_{x \rightarrow p} \frac{\sqrt{D(p, x)}}{\|x - p\|^{1-\beta}} < \infty \right\} \quad (7.19)$$

and we say that h is *tight* at p if the infimum is attained in (7.19), i.e., if $\beta_h(p)$ is the minimal $\beta \in [0, 1]$ such that

$$D(p, x) = \mathcal{O}(\|p - x\|^{2(1-\beta)}) \quad \text{for } x \text{ near } p. \quad (7.20)$$

Informally, the Legendre exponent measures the deficit in relative size between ordinary “norm neighborhoods” in $\mathcal{X}$ and the corresponding “Bregman neighborhoods” induced by the sublevel sets of the Bregman divergence. Specifically, (i) the case $\beta_h(p) = 0$ corresponds to the “norm-like” behavior $D(p, x) = \Theta(\|p - x\|^2)$; (ii) any other value $\beta_h(p) \in (0, 1)$ indicates a different limiting behavior for $D(p, x)$ as $x \rightarrow p$; and, finally, (iii) when $\beta_h(p) = 1$ we may have $\limsup_{x \rightarrow p} D(p, x) > 0$. In this last case, the ambient norm topology is *strictly coarser* than the Bregman topology in the sense that $D(p, x_t)$ may remain bounded away from zero even if $x_t \rightarrow p$ as $t \rightarrow \infty$. The following remark illustrates this phenomenon with the Hellinger regularize.

Remark 10. For example, let $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ be the unit Euclidean ball in $\mathbb{R}^n$ and consider the n -dimensional Hellinger regularizer $h(x) = -\sqrt{1 - \|x\|_2^2}$. Then, for all p on the boundary of $\mathcal{X}$ and all $x \in \mathcal{X}_h = \text{int } \mathcal{X}$, we readily get

$$D(p, x) = \frac{1 - \sum_{i=1}^n p_i x_i}{\sqrt{1 - \|x\|_2^2}} = \frac{1 - \langle p, x \rangle}{\sqrt{1 - \|x\|_2^2}}. \quad (7.21)$$

Shadowing the Euclidean definition, a “Hellinger sphere” of radius r centered at p is defined as the corresponding level set of $D(p, x)$, i.e., $S_p(r) = \{x \in \mathcal{X}_h : D(p, x) = r\}$. Hence, by (7.21), this means that a point $x \in \mathcal{X}_h$ is at “Hellinger distance” r relative to p if

$$1 - \sum_i p_i x_i = r \sqrt{1 - \|x\|_2^2}. \quad (7.22)$$

Now, if we take a sequence x_t converging to p in the ordinary Euclidean sense (i.e., $\|x_t - p\|_2 \rightarrow 0$), both sides of (7.22) converge to 0. This has several counterintuitive consequences: (i) the “Hellinger ball” $\mathcal{B}_r^h(p) := \{x \in \mathcal{X}_h : D(p, x) \leq r^2/2\}$ is *not closed* in the Euclidean topology; and (ii) the “Hellinger center” p of $\mathcal{B}_r^h(p)$ actually belongs to the Euclidean boundary of $\mathcal{B}_r^h(p)$. As a result, for all $n \geq 2$, it is straightforward to construct a sequence x_t with $\|x_t - p\|_2 \rightarrow 0$, but which remains at *constant* Hellinger divergence relative to p .³ 

For illustration purposes, we compute below the Legendre exponent for each of the running examples of § 7.3 (see also Table 1):

1. *Quadratic regularization* (Example 7.1): Since $D(p, x) = (p - x)^2/2$ for all $p, x \in \mathcal{X}$, we have $\beta_h(p) = 0$ for all $p \in \mathcal{X}$.
2. *Negative entropy* (Example 7.2): For $p = 0$, (7.10) gives $D(0, x) = x$, so $\beta_h(0) = 1/2$. Otherwise, for all $p \in \mathcal{X}_h = (0, \infty)$, a Taylor expansion with Lagrange remainder yields $D(p, x) = \mathcal{O}((p - x)^2)$, so $\beta_h(p) = 0$ for all $p \in (0, \infty)$.
3. *Tsallis entropy* (Example 7.3): For $p = 0$, (7.13) gives $D(0, x) = x^q/q$, so $\beta_h(0) = \max\{0, 1 - q/2\}$. Otherwise, for all $p \in \mathcal{X}_h = (0, \infty)$, a Taylor expansion yields $D(p, x) = \mathcal{O}((p - x)^2)$, so $\beta_h(p) = 0$ in this case.
4. *Hellinger regularizer* (Example 7.4): For $p = \pm 1$, (7.15) readily gives $D(\pm 1, x) = \sqrt{(1 \mp x)/(1 \pm x)} = \Theta(|x \mp 1|^{1/2})$, so $\beta_h(\pm 1) = 1 - 1/4 = 3/4$. Instead, if $p \in (-1, 1)$ a Taylor expansion again yields $D(p, x) = \mathcal{O}((p - x)^2)$, so $\beta_h(p) = 0$ in this case.

A common pattern that emerges above is that $\beta_h(p) = 0$ whenever p is an interior point. We make this observation precise in Lemma E.3.1 in Appendix E, where we show more generally that $\beta_h(p) = 0$ whenever ∇h is (locally) Lipschitz continuous in a neighborhood of p in $\mathcal{X}$.

7.4.2 Convergence rate analysis

We are now in a position to state our first general result for the convergence rate of (BPM). To do so, we will make the blanket assumption that h is tight at x^* with Legendre exponent $\beta^* := \beta_h(x^*)$. In particular, this means that there exists a neighborhood $\mathcal{U}$ of x^* in $\mathcal{X}$ and a positive constant $K > 0$ such that

$$D(x^*, x) \leq \frac{K}{2} \|x - x^*\|^{2(1-\beta^*)} \quad \text{for all } x \in \mathcal{U}. \quad (7.23)$$

To lighten notation, we will also assume that h is 1-strongly convex on $\mathcal{U}$ (cf. Remark 9 and the beginning of § 7.4.1). We then have the following result.

Theorem 7.2. *Suppose that Assumptions 7.1, 7.2, and 7.4 hold and (BPM) is run with a constant step-size $\gamma_t \equiv \gamma$, $t = 1, 2, \dots$, such that*

$$\gamma \leq \frac{1}{2\varphi L} \quad \text{and} \quad \gamma(1 - a - b)^2 \leq \frac{\mu}{8L^2} \quad (7.24)$$

³ For instance, if $n = 2$, the point $x_u = (1 - u, \sqrt{2u(1 - u)})$ converges to $p = (1, 0)$ as $u \rightarrow 0^+$, even though $D(p, x_u) = 1$ for all $u \in (0, 1)$. Crucially, if $n = 1$, this phenomenon does not occur (cf. Example 7.4).

	Domain ($\mathcal{X}$)	Regularizer (h)	Legendre Exponent ($\beta_h(p)$)	Convergence Rate
EUCLIDEAN	arbitrary	$x^2/2$	0	Linear
ENTROPIC	$[0, \infty)$	$x \log x$	$1/2$	$\mathcal{O}(1/t)$
TSALLIS	$[0, \infty)$	$[q(1-q)]^{-1}(x - x^q)$	$\max\{0, 1 - q/2\}$	$\mathcal{O}(1/t^{q/(2-q)})$
HELLINGER	$[-1, 1]$	$-\sqrt{1-x^2}$	$3/4$	$\mathcal{O}(1/t^{1/3})$

Table 1: Summary of the Legendre exponents for the 1-dimensional examples of § 7.2 at a boundary point p of $\mathcal{X} \subset \mathbb{R}$, and the associated convergence rates in terms of the Bregman divergence $D(x^*, x_t)$.

where $\varphi = (\sqrt{5} + 1)/2$ is the golden ratio. If x_1 is initialized sufficiently close to x^* , the iterates x_t of (BPM) enjoys the bound

$$D(x^*, x_t) \leq D(x^*, x_1) \cdot \begin{cases} (1 - \frac{\mu\gamma}{2K})^{t-1} & \text{if } \beta^* = 0, \\ [1 + \rho\mu\gamma(t-1)]^{1-1/\beta^*} & \text{if } \beta^* \in (0, 1), \end{cases} \quad (7.25)$$

where $\rho = \frac{\beta^*}{1-\beta^*} \max\{2K^{\frac{1}{1-\beta^*}} D(x^*, x_1)^{-\frac{\beta^*}{1-\beta^*}}, 2^{\frac{\beta^*}{1-\beta^*}}\}^{-1}$.

Before moving on to the proof of Theorem 7.2, some remarks and corollaries are in order (see also Table 1 for an explicit illustration of the derived rates for Examples 7.1–7.4):

Remark 1. The first point of note is the sharp drop in the convergence rate of (BPM) from geometric, when $\beta^* = 0$, to a power law when $\beta^* > 0$. As we saw in § 7.3, this drop is unavoidable, even when $\mathcal{X}$ is 1-dimensional and g is affine; in fact, the calculations of § 7.3 show that the rates provided by Theorem 7.2 are, in general, unimprovable. 🍷

Remark 2. We should also note that the guarantees of Theorem 7.2 are stated in terms of the Bregman divergence, not the ambient norm. Since $D(x^*, x_t) = \Omega(\|x_t - x^*\|^2)$, these bounds can be restated in terms of $\|x_t - x^*\|$, but this conversion is not without loss of information: if the bound $D(x^*, x_t) = \Omega(\|x_t - x^*\|^2)$ is not tight, the actual rate in terms of the norm may be significantly different. This phenomenon was already observed in the 1-dimensional examples of § 7.3 where $D(x^*, x_t) = \Theta(\|x_t - x^*\|^{2(1-\beta^*)})$, in which case Theorem 7.2 gives

$$\|x_t - x^*\| = \mathcal{O}(t^{-1/(2\beta^*)}) \quad (7.26)$$

whenever $\beta^* > 0$ (see also Table 1). In general however, the Bregman divergence may grow at different rates along different rays emanating from x^* , so it is not always possible to translate a Bregman-based bound to a norm-based bound (or vice versa). This analysis requires a much closer look at the geometric structure of $\mathcal{X}$, depending on which constraints are active at x^* ; we examine this issue at depth in § 7.5. 🍷

Remark 3. We should also note that, even though Theorem 7.2 is stated for a constant step-size, our proof allows for a variable step-size γ_t , provided that the step-size conditions (7.24) are satisfied. In this case, the bounds (7.25) becomes

$$D(x^*, x_t) \leq D(x^*, x_1) \cdot \begin{cases} \prod_{s=1}^{t-1} (1 - \frac{\mu\gamma_s}{2K}) & \text{if } \beta^* = 0, \\ [1 + \rho\mu \sum_{s=1}^{t-1} \gamma_s]^{1-1/\beta^*} & \text{if } \beta^* \in (0, 1). \end{cases}$$

7.4.3 Proof of [Theorem 7.2](#)

We now proceed to the proof of [Theorem 7.2](#), beginning with a series of intermediate results tailored to the update structure of (BPM). The first of these lemmas relates the Bregman divergence before and after a prox-step modulo an element of the polar cone $\text{PC}(p) := \{v \in \mathcal{Y} : \langle v, x - p \rangle \leq 0 \text{ for all } x \in \mathcal{X}\}$ of $\mathcal{X}$ at the reference point p .

Lemma 7.3. *Let $x^+ = P_x(y)$ for some $x \in \mathcal{U} \cap \mathcal{X}_h$, $y \in \mathcal{Y}$, such that $x^+ \in \mathcal{U}$. Then, for all $p \in \mathcal{X}$ and all $v \in \text{PC}(p)$, we have:*

$$D(p, x^+) \leq D(p, x) + \langle y - v, x^+ - p \rangle - D(x^+, x) \quad (7.27a)$$

$$\leq D(p, x) + \langle y - v, x - p \rangle + \frac{1}{2} \|y - v\|_*^2, \quad (7.27b)$$

The next lemma extends [Lemma 7.3](#) to emulate the two-step structure of (BPM):

Lemma 7.4. *Let $x_i^+ = P_x(y_i)$ for some $x \in \mathcal{U} \cap \mathcal{X}_h$, $y_i \in \mathcal{Y}$, such that $x_i^+ \in \mathcal{U}$, $i = 1, 2$. Then, for all $p \in \mathcal{X}$ and all $v \in \text{PC}(p)$, we have:*

$$D(p, x_2^+) \leq D(p, x) + \langle y_2 - v, x_1^+ - p \rangle + \frac{1}{2} \|y_2 - y_1 - v\|_*^2 - \frac{1}{2} \|x_1^+ - x\|^2. \quad (7.28)$$

Versions of the above inequalities already exist in the literature, see e.g., ([Juditsky et al., 2011](#), Lem. 4), ([Mertikopoulos et al., 2019](#), Prop. B.4). The novelty in [Lemmas 7.3](#) and [7.4](#) is the extra term involving the polar vector $v \in \text{PC}(p)$; this term plays an important role in the sequel, so we provide a complete proof in [Appendix E](#).

With these preliminaries in hand, we proceed to derive two further inequalities that play a pivotal role in the analysis of (BPM). The first is an immediate corollary of [Lemma 7.4](#):

Corollary 7.5. *Let x^* be a solution of (VI) and suppose that $x_t, x_{t+1/2} \in \mathcal{U}$ for all t . Then, for all $c \geq 0$ and all $t = 1, 2, \dots$, the iterates of (BPM) satisfy the template inequality*

$$\begin{aligned} D(x^*, x_{t+1}) &\leq D(x^*, x_t) - \gamma_t \langle g_{t+1/2} - c g(x^*), x_{t+1/2} - x^* \rangle \\ &\quad + \frac{1}{2} \gamma_t^2 \|g_{t+1/2} - g_t - c g(x^*)\|_*^2 - \frac{1}{2} \|x_{t+1/2} - x_t\|^2. \end{aligned} \quad (7.29)$$

Proof. Since x^* is a solution of (VI), we have $g(x^*) \in -\text{PC}(x^*)$. (7.29) then follows from [Lemma 7.4](#) with $x \leftarrow x_t$, $p \leftarrow x^*$, $v \leftarrow -c \gamma_t g(x^*) \in \text{PC}(x^*)$ and $(y_1, y_2) \leftarrow (-\gamma_t g_t, -\gamma_t g_{t+1/2})$. $\blacksquare$

The second inequality that we derive provides an “energy function” for (BPM), namely

$$E_t = D_t + f_t \quad (7.30)$$

where $D_t = D(x^*, x_t)$ and $f_t = \gamma_{t-1}^2 \|(a+b)g_{t-1/2} - g_{t-1}\|_*^2$ (by convention, we take $f_1 = 0$). The lemma below outlines the Lyapunov properties of E_t .

Proposition 7.6. *Suppose that [Assumptions 7.1](#) and [7.4](#) hold and (BPM) is run with a step-size such that*

$$\lambda \gamma_t + 4 \gamma_t^2 L^2 \leq 1 \quad \text{for some } \lambda \geq 0 \text{ and all } t = 1, 2, \dots \quad (7.31)$$

Suppose further that $x_t, x_{t+1/2} \in \mathcal{U}$ for all t . Then the iterates x_t of (BPM) satisfy the inequality

$$\begin{aligned} E_{t+1} &\leq E_t - \lambda \gamma_t f_t - \gamma_t \langle g(x_{t+1/2}) - g(x^*), x_{t+1/2} - x^* \rangle \\ &\quad - \gamma_t(a+b) \langle g(x^*), x_{t+1/2} - x^* \rangle - \frac{1}{2} \|x_{t+1/2} - x_t\|^2 \\ &\quad + \gamma_t^2(1-a-b)^2 L^2 \|x_{t+1/2} - x^*\|^2 \\ &\quad + 2\gamma_t^2(a+b)^2 L^2 \|x_{t+1/2} - x_t\|^2. \end{aligned} \quad (7.32)$$

Proof. Let $c = 1 - a - b$ so $c \geq 0$ by Assumption 7.4. Corollary 7.5 then yields

$$\begin{aligned} D_{t+1} &\leq D_t - \gamma_t \langle g_{t+1/2} - g(x^*), x_{t+1/2} - x^* \rangle \\ &\quad - \gamma_t(a+b) \langle g(x^*), x_{t+1/2} - x^* \rangle - \frac{1}{2} \|x_{t+1/2} - x_t\|^2 \\ &\quad + \frac{\gamma_t^2}{2} \|g_{t+1/2} - g_t - c g(x^*)\|_*^2. \end{aligned} \quad (7.33)$$

Since $g_{t+1/2} = (a+b)g_{t+1/2} + c g_{t+1/2}$, the last term above may be bounded as

$$\begin{aligned} \frac{1}{2} \gamma_t^2 \|g_{t+1/2} - g_t - c g(x^*)\|_*^2 &\leq \gamma_t^2 c^2 \|g_{t+1/2} - g(x^*)\|_*^2 + \gamma_t^2 \|(a+b)g_{t+1/2} - g_t\|_*^2 \\ &\leq \gamma_t^2 c^2 L^2 \|x_{t+1/2} - x^*\|_*^2 + \gamma_t^2 \|(a+b)g_{t+1/2} - g_t\|_*^2 \\ &= \gamma_t^2(1-a-b)^2 L^2 \|x_{t+1/2} - x^*\|^2 + f_{t+1}, \end{aligned} \quad (7.34)$$

where we used Assumption 7.1 in the second line and the definition (7.30) of f_t in the last one. Thus, combining (7.33) and (7.34) and comparing to (7.32), it suffices to show that

$$2f_{t+1} \leq (1 - \lambda \gamma_t) f_t + 4\gamma_t^2(a+b)^2 L^2 \|x_{t+1/2} - x_t\|^2 \quad \text{for all } t = 1, 2, \dots \quad (7.35)$$

We consider two distinct cases for this below.

CASE 1: $t = 1$ By the definition (7.30) of f_t and (7.4) and (7.5), we have:

$$\begin{aligned} f_2 &= \gamma_1^2 \|(a+b)g_{3/2} - g_1\|_*^2 = \gamma_1^2(a+b)^2 \|g(x_{3/2}) - g(x_1)\|_*^2 \\ &\leq \gamma_1^2(a+b)^2 L^2 \|x_{3/2} - x_1\|^2, \end{aligned}$$

where we used the initialization assumption $x_1 = x_{1/2}$ in the second equality and the Lipschitz continuity of g in the last one. Since $f_1 = 0$ by construction, our claim is immediate.

CASE 2: $t > 1$ By Young's inequality and the Lipschitz continuity of g , we readily obtain

$$\begin{aligned} f_{t+1} &= \gamma_t^2 \|(a+b)g_{t+1/2} - g_t\|_*^2 \\ &= \gamma_t^2 \|(a+b)[g(x_{t+1/2}) - g(x_t)] + b[g(x_t) - g(x_{t-1/2})]\|_*^2 \\ &\leq 2\gamma_t^2(a+b)^2 \|g(x_{t+1/2}) - g(x_t)\|_*^2 + 2\gamma_t^2 b^2 \|g(x_t) - g(x_{t-1/2})\|_*^2 \\ &\leq 2\gamma_t^2(a+b)^2 L^2 \|x_{t+1/2} - x_t\|^2 + 2\gamma_t^2 b^2 L^2 \|x_t - x_{t-1/2}\|^2 \\ &\leq 2\gamma_t^2(a+b)^2 L^2 \|x_{t+1/2} - x_t\|^2 + 2\gamma_t^2 b^2 L^2 \gamma_{t-1}^2 \|g_{t-1/2} - g_{t-1}\|_*^2 \end{aligned}$$

where, in the last line, we used [Lemma E.2.3](#) to bound the difference $x_t - x_{t-1/2}$ as

$$\|x_t - x_{t-1/2}\| = \|P_{x_{t-1}}(-\gamma_{t-1}g_{t-1/2}) - P_{x_{t-1}}(-\gamma_{t-1}g_{t-1})\| \leq \gamma_{t-1}\|g_{t-1/2} - g_{t-1}\|_*. \quad (7.36)$$

Finally, by [Assumption 7.4](#), we have $c = 0$ whenever $b > 0$, so $b^2\gamma_{t-1}^2\|g_{t-1/2} - g_{t-1}\|_*^2 = b^2\gamma_{t-1}^2\|(1-c)g_{t-1/2} - g_{t-1}\|_*^2 = b^2f_t$ for all $t > 1$. Hence, putting everything together, we get

$$f_{t+1} \leq 2\gamma_t^2(a+b)^2L^2\|x_{t+1/2} - x_t\|^2 + 2\gamma_t^2b^2L^2f_t. \quad (7.37)$$

(7.35) then follows by the requirement (7.31), which implies that $2\gamma_t^2L^2 \leq (1 - \lambda\gamma_t)/2$. $\blacksquare$

Moving forward, since x^* is a solution of (VI), the first line of (7.32) yields a negative $\mathcal{O}(\gamma_t)$ contribution to E_t , whereas the third and fourth lines collectively represent a subleading $\mathcal{O}(\gamma_t^2)$ “error term”. This decomposition would suffice for the analysis of (BPM) if the coupling term $\langle g(x^*), x_{t+1/2} - x^* \rangle$ did not incur an additional $\mathcal{O}(\gamma_t)$ positive contribution to E_{t+1} . This error term is difficult to control but if x^* satisfies (SOS), we have the following bound.

Lemma 7.7. *Suppose that [Assumption 7.2](#) holds. Then, for all $x \in \mathcal{X}$, $x' \in \mathcal{B}$ and all $c \in [0, 1]$, we have:*

$$\langle g(x') - cg(x^*), x' - x^* \rangle \geq \frac{1}{2}\mu\|x - x^*\|^2 - \mu\|x' - x\|^2. \quad (7.38)$$

Proof. Since x^* is a solution of (VI) and $c \in [0, 1]$, we have $(1-c)\langle g(x^*), x' - x^* \rangle \geq 0$ for all $x' \in \mathcal{X}$. Hence, by [Assumption 7.2](#), we get

$$\langle g(x') - cg(x^*), x' - x^* \rangle \geq \langle g(x') - g(x^*), x' - x^* \rangle \geq \mu\|x' - x^*\|^2 \quad (7.39)$$

and our assertion follows from the basic bound $\|x - x^*\|^2 \leq 2\|x - x'\|^2 + 2\|x' - x^*\|^2$. $\blacksquare$

With this ancillary estimate in hand, we may finally sharpen [Proposition 7.6](#) to obtain a bona fide energy inequality for solutions satisfying (SOS):

Proposition 7.8. *Suppose that [Assumptions 7.1](#), [7.2](#), and [7.4](#) hold and (BPM) is run with γ_t such that*

$$2\mu\gamma_t + 4\gamma_t^2L^2 \leq 1 \quad \text{and} \quad (1-a-b)^2\gamma_t \leq \frac{\mu}{8L^2} \quad \text{for all } t = 1, 2, \dots \quad (7.40)$$

Suppose further that $x_t, x_{t+1/2} \in \mathcal{U}$ for all t . Then, for all $t \geq 1$ such that $x_{t+1/2} \in \mathcal{B}$, we have

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \frac{1}{4}\mu\gamma_t\|x_t - x^*\|^2. \quad (7.41)$$

Proof. Assume that $x_{t+1/2} \in \mathcal{B}$ and set $c = 1 - a - b$. Then, invoking [Lemma 7.7](#) with $x \leftarrow x_t$ and $x' \leftarrow x_{t+1/2}$, we get

$$\begin{aligned} & \langle g(x_{t+1/2}) - g(x^*), x_{t+1/2} - x^* \rangle + (a+b)\langle g(x^*), x_{t+1/2} - x^* \rangle \\ & \geq \frac{1}{2}\mu\|x_t - x^*\|^2 - \mu\|x_{t+1/2} - x_t\|^2. \end{aligned} \quad (7.42)$$

Thus, taking $\lambda \leftarrow \mu$ in [Proposition 7.6](#) (in terms of step-size conditions, the first part of (7.40) implies (7.31)) and combining with the above, the bound (7.32) becomes

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \frac{1}{2}\mu\gamma_t\|x_t - x^*\|^2 + \gamma_t^2c^2L^2\|x_{t+1/2} - x^*\|^2$$

$$-\frac{1}{2}(1 - 4\gamma_t^2(a+b)^2L^2 - 2\mu\gamma_t)\|x_{t+1/2} - x_t\|^2. \quad (7.43)$$

Hence, writing $\|x_{t+1/2} - x^*\|^2 \leq 2\|x_{t+1/2} - x_t\|^2 + 2\|x_t - x^*\|^2$ and rearranging, we obtain

$$\begin{aligned} E_{t+1} &\leq E_t - \mu\gamma_t f_t - \frac{1}{2}(\mu\gamma_t - 4\gamma_t^2c^2L^2)\|x_t - x^*\|^2 \\ &\quad - \frac{1}{2}(1 - 4\gamma_t^2((a+b)^2 + c^2)L^2 - 2\mu\gamma_t)\|x_{t+1/2} - x_t\|^2. \end{aligned} \quad (7.44)$$

Since $a, b, c \geq 0$ and $a + b + c = 1$, we also have $(a+b)^2 + c^2 \leq 1$, so the step-size assumption (7.40) guarantees that the last term in (7.44) is nonpositive. Likewise, the second part of (7.40) gives $\mu\gamma_t - 4\gamma_t^2c^2L^2 \geq \frac{1}{2}\mu\gamma_t$, so the energy inequality (7.41) follows and our proof is complete. ■

We finally have all the required building blocks in place to prove [Theorem 7.2](#).

Proof of Theorem 7.2. Our proof strategy consists of the following basic steps:

1. We first show that, if the step-size of (BPM) satisfies (7.24) and x_1 is initialized sufficiently close to x^* , the base and leading state sequences x_t and $x_{t+1/2}$, $t = 1, 2, \dots$, both remain within the neighborhood $\mathcal{U} \cap \mathcal{B}$ of x^* where (SOS), (7.18), and (7.23) all hold.
2. By virtue of this stability result, the energy inequality (7.41) and the definition of the Legendre exponent allow us to express $D_t = D(x^*, x_t)$ as $D_{t+1} \leq D_t - \mathcal{O}(D_t^{1/(1-\beta^*)})$ up to an error term that vanishes at a geometric rate. The rates (7.25) are then derived by analyzing this recursive inequality for $\beta^* = 0$ and $\beta^* > 0$ respectively.

We now proceed to detail the two steps outlined above.

STEP 1: STABILITY Take $r > 0$ such that $\mathcal{B}_r^{\mathcal{X}}(x^*) := \{x \in \mathcal{X} : \|x - x^*\| \leq r\} \subset \mathcal{B} \cap \mathcal{U}$ and, such that, for any $x \in \mathcal{B}_r^{\mathcal{X}}(x^*) \cap \mathcal{X}_h$, $x' \in \mathcal{B}_r^{\mathcal{X}}(x^*)$, $P_x(-\gamma g(x'))$ still belongs to $\mathcal{U}$, for all admissible step-sizes γ . This is indeed possible by the continuity of the prox-mapping (see [Lemma E.2.3](#) in [Appendix E](#)) and of g . Assume further that $x_{1/2} = x_1 \in \mathcal{B}_r^{\mathcal{X}}(x^*)$ is such that $D(x^*, x_{1/2}) = D(x^*, x_1) \leq (1 - \lambda)r^2/4$, where $\lambda \in (0, 1)$ is a constant to be determined later. Then, letting $E_0 := D(x^*, x_1)$, we will prove by induction on t that

$$\max\{\|x_{t-1/2} - x^*\|, \|x_t - x^*\|\} \leq r \quad \text{and} \quad E_t \leq E_{t-1}, \quad (7.45)$$

which will show in particular that $x_{t+1/2} \in \mathcal{B} \cap \mathcal{U}$ for all $t \geq 1$. Indeed:

- For the base case ($t = 1$), we have $x_{1/2} = x_1 \in \mathcal{B}_r^{\mathcal{X}}(x^*)$ and $E_1 = E_0$ by construction, so there is nothing to show.
- For the induction step, assume (7.45) holds. Then, since $x_t \in \mathcal{B}_r^{\mathcal{X}}(x^*) \subset \mathcal{U}$, (7.18) yields

$$\frac{1}{2}\|x_t - x^*\|^2 \leq D_t \leq E_t \leq E_1 = D(x^*, x_1) \quad (7.46)$$

Moreover, both x_t and $x_{t-1/2}$ are in $\mathcal{B}_r^{\mathcal{X}}(x^*)$ so that, by construction, $x_{t+1/2}$ is still in $\mathcal{U}$. Now, to show that $x_{t+1/2} \in \mathcal{B}_r^{\mathcal{X}}(x^*)$, [Lemma 7.3](#) with $p \leftarrow x^*$, $x \leftarrow x_t$, $y \leftarrow -\gamma_t g_t$ and $v \leftarrow -(a+b)\gamma_t g(x^*)$ gives

$$\begin{aligned} D_{t+1/2} &\leq D_t - \gamma_t \langle g_t - (a+b)g(x^*), x_{t+1/2} - x^* \rangle \\ &\leq D_t - a\gamma_t \langle g(x_t) - g(x^*), x_{t+1/2} - x^* \rangle \\ &\quad - b\gamma_t \langle g(x_{t-1/2}) - g(x^*), x_{t+1/2} - x^* \rangle \end{aligned} \quad (7.47)$$

and hence, by Young's inequality and (7.18), we get

$$\begin{aligned} \frac{1}{2} \|x_{t+1/2} - x^*\|^2 &\leq D_t + \gamma_t^2 a \|g(x_t) - g(x^*)\|_*^2 + \gamma_t^2 b \|g(x_{t-1/2}) - g(x^*)\|_*^2 \\ &\quad + \frac{1}{4}(a+b) \|x_{t+1/2} - x^*\|^2. \end{aligned} \quad (7.48)$$

Since $a + b \leq 1$, using Assumption 7.1 and rearranging gives

$$\begin{aligned} \|x_{t+1/2} - x^*\|^2 &\leq 4D_t + 4\gamma_t^2 L^2 \max\{\|x_t - x^*\|^2, \|x_{t-1/2} - x^*\|^2\} \\ &\leq (1-\lambda)r^2 + 4\gamma_t^2 L^2 r^2 \end{aligned} \quad (7.49)$$

where we used the fact that $\|x_{t-1/2} - x^*\|^2 \leq r^2$ and $\|x_t - x^*\|^2 \leq 2D_t \leq \frac{1}{2}(1-\lambda)r^2$ (by the inductive hypothesis and (7.46) respectively). Thus, with $2\gamma_t L \leq 1/\varphi < 1$ by assumption, choosing $\lambda = 1/\varphi^2$ gives $\|x_{t+1/2} - x^*\|^2 \leq r^2$, which completes the first part of the induction. Finally, for the second part, we have $x_{t+1} \in \mathcal{U}$ because $x_t, x_{t+1/2}$ have been shown to be in $\mathcal{B}_r^{\mathcal{X}}(x^*)$ and our step-size assumption gives

$$2\mu\gamma_t + 4\gamma_t^2 L^2 \leq 2\gamma_t L + 4\gamma_t^2 L^2 \leq 1/\varphi + 1/\varphi^2 = 1. \quad (7.50)$$

Thus, since $x_{t+1/2} \in \mathcal{B}$, Proposition 7.8 readily gives

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \frac{1}{4}\mu\gamma_t \|x_t - x^*\|^2 \leq E_t, \quad (7.51)$$

and the induction is complete.

STEP 2: CONVERGENCE RATE ANALYSIS From (7.51) and the Legendre bound (7.23), we get

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \frac{\mu\gamma_t}{2^{1-\alpha}K^{1+\alpha}} D_t^{1+\alpha} \quad \text{with } \alpha = \beta^*/(1-\beta^*). \quad (7.52)$$

We now distinguish two cases, depending on whether $\beta^* = 0$ or $\beta^* > 0$.

Case 1: If $\beta^* = 0$, we have $\alpha = 0$ by definition and $K \geq 1$ by (7.18). (7.52) then gives

$$E_{t+1} \leq E_t - \frac{\mu\gamma_t}{2K} D_t - \mu\gamma_t f_t \leq \left(1 - \frac{\mu\gamma_t}{2K}\right) E_t \quad (7.53)$$

so the case $\beta = 0$ of (7.25) follows immediately by setting $\gamma_t \equiv \gamma$ for all t .

Case 2: If $\beta^* > 0$, then $\alpha > 0$ too, so we will proceed by rewriting all terms in (7.52) in terms of E_t . To that end, we have:

$$\begin{aligned} E_{t+1} &\leq E_t - \mu\gamma_t f_t - \frac{\mu\gamma_t}{2^{1-\alpha}K^{1+\alpha}} D_t^{1+\alpha} \\ &\leq E_t - \frac{\mu\gamma_t}{D(x^*, x_1)^\alpha} f_t^{1+\alpha} - \frac{\mu\gamma_t}{2^{1-\alpha}K^{1+\alpha}} D_t^{1+\alpha} \\ &\leq E_t - \frac{\mu\gamma_t}{\max(2^{1-\alpha}K^{1+\alpha}, D(x^*, x_1)^\alpha)} [D(x^*, x_t)^{1+\alpha} + f_t^{1+\alpha}] \\ &\leq E_t - \frac{\mu\gamma_t}{\max(2K^{1+\alpha}, 2^\alpha D(x^*, x_1)^\alpha)} E_t^{1+\alpha} \end{aligned} \quad (7.54)$$

where, in the second line, we used (7.45) to get $f_t \leq D(x^*, x_t) + f_t \leq D(x^*, x_1)$, and, in the last line, we used the convexity of $x^{1+\alpha}$. The case $\beta \in (0, 1)$ of (7.25) then follows from Lemma 6 of (Polyak, 1987, p. 46) (recreated as Lemma E.1.1 in Appendix E).

■

7.5 FINER RESULTS FOR LINEARLY CONSTRAINED PROBLEMS

Our goal in this last section is to take a closer look at the convergence rate of (BPM) for different solution configurations that arise in linearly constrained problems. To that end, we begin by revisiting the examples of § 7.3.

7.5.1 Motivating examples, redux

A common feature of Examples 7.1–7.4 is that the problem’s defining vector field vanishes at the solution point under study. In the series of examples below, we examine the rate of convergence achieved when this is not the case.

Example 7.5 (Euclidean regularization). Consider again the quadratic regularizer of Example 7.1 over $\mathcal{X} = [0, \infty)$, but with $g(x) = x + 1$. The solution of (VI) is still $x^* = 0$ but the update (7.7) now becomes

$$F(x) = [x - \gamma(x + 1)]_+ = [(1 - \gamma)x - \gamma]_+. \quad (7.55)$$

Since $F(x) = 0$ for all sufficiently small $x > 0$, we readily conclude that x_t converges to x^* in a *finite* number of iterations. 🐼

Example 7.6 (Entropic regularization). Under the entropic regularizer of Example 7.2, and taking again $g(x) = x + 1$, the update rule (7.7) becomes

$$F(x) = x \exp(-\gamma(x + 1)) = xe^{-\gamma} + o(x) \sim xe^{-\gamma} \quad (7.56)$$

i.e., F is a contraction for small $x > 0$. Hence, in contrast to Example 7.2, x_t converges to 0 at a geometric rate, even though the problem’s solution lies on the boundary of $\mathcal{X}$. 🐼

Example 7.7 (Fractional power). Finally, consider the fractional power regularizer of Example 7.3, again with $g(x) = x + 1$. Then, for $q \in (0, 1)$, the update rule (7.7) gives

$$F(x) = [x^{q-1} + \gamma(1 - q)(x + 1)]^{1/(q-1)} = x - \gamma x^{2-q} + o(x^{2-q}) \quad (7.57)$$

for small $x > 0$. Thus, by Lemma 7.1, we get that x_t converges to 0 as $|x_t - x^*| = \Theta(t^{-1/(1-q)})$ and $D(x^*, x_t) = \Theta(t^{-q/(1-q)})$, which is again faster than the rate given by Theorem 7.2. 🐼

Examples 7.5–7.7 show that the convergence rate of (BPM) when x^* in the boundary can change drastically depending on whether $g(x^*)$ is zero or not. In the example below, we examine in more detail the behavior of the individual coordinates of x_t as a function of the position of $g(x^*)$ relative to $\mathcal{X}$.

Example 7.8 (Higher-dimensional simplices). Consider the canonical two-dimensional simplex $\mathcal{X} = \{(x_1, x_2, x_3) \in \mathbb{R}_+^3 : x_1 + x_2 + x_3 = 1\}$ of $\mathbb{R}^3$ equipped with the entropic regularizer $h(x) = \sum_{i=1}^3 x_i \log x_i$. Consider also the vector field $g(x) = x - p$ with $p = (-v_1, -v_2, 1)$ for some $v_1, v_2 \geq 0$, so the solution of (VI) is $x^* = (0, 0, 1)$, an extreme point of $\mathcal{X}$.

Since the Legendre exponent of h at x^* is easily seen to be $\beta_h(x^*) = 1/2$, Theorem 7.2 would indicate a rate of convergence of $D(x^*, x_t) = \mathcal{O}(1/t)$ or, in terms of norms, $\|x_t - x^*\| = \mathcal{O}(1/t)$. However, this rate can be very pessimistic

if, for example, $v_1 > 0$. Indeed, in this case, since x_t converges to $x^* = (0, 0, 1)$, the relevant coordinates of $g(x_t)$ will evolve as $g_1(x_t) = x_{1,t} + v_1 = v_1 + o(1)$ and $g_3(x_t) = x_{3,t} - 1 = o(1)$. Accordingly, since entropic regularization on the simplex leads to the exponential weights update [Beck and Teboulle \(2003\)](#)

$$x_{i,t+1} \propto x_{i,t} \exp(-\gamma g_i(x_t)) \quad \text{for all } t \geq 1, i = 1, 2, 3, \quad (7.58)$$

the fact that $\lim_{t \rightarrow \infty} x_{3,t} = 1$ readily yields

$$x_{1,t+1} \sim \frac{x_{1,t+1}}{x_{3,t+1}} = \frac{x_{1,t}}{x_{3,t}} \exp(-\gamma g_1(x_t) + \gamma g_3(x_t)) = \frac{x_{1,t}}{x_{3,t}} \exp(-\gamma v_1 + o(1)) \quad (7.59)$$

i.e., $x_{1,t}$ converges to 0 at a *geometric* rate whenever $v_1 > 0$.

By symmetry, the argument above yields the same rate for $x_{2,t}$ if $v_2 > 0$. However, as we show in [Appendix E.5](#), if $v_2 = 0$, we would have $x_{2,t} = \Theta(1/t)$ no matter the value of v_1 (and likewise for the rate of $x_{1,t}$ if $v_1 = 0$). In other words, the rate provided by [Theorem 7.2](#) is tight for the coordinate $i \in \{1, 2\}$ with a vanishing drift coefficient v_i , but not otherwise; we will devote the rest of this section to deriving a formal statement (and proof) of the general principle underlying this observation. 

7.5.2 Linearly constrained problems

For concreteness, we focus below on linearly constrained problems where the different convergence behaviors outlined in the previous examples can be characterized in a precise manner. To lighten notation, we identify $\mathcal{V}$ with $\mathbb{R}^n$ endowed with the Euclidean scalar product $\langle \cdot, \cdot \rangle$, and we will not distinguish between primal and dual vectors (so the distinction between normal and polar cones is likewise blurred).

Formally, we consider polyhedral domains written in standard form as

$$\mathcal{X} = \{x \in \mathbb{R}_+^n : Ax = b\} \quad (7.60)$$

for some matrix $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.⁴ Moreover, to avoid trivialities, we further assume that $\mathcal{X}$ admits a *Slater point*, i.e., there exists some $x \in \mathcal{X}$ such that $x_i > 0$ for all $i = 1, \dots, n$. This setup is particularly flexible, as it allows us to identify the *active* constraints at $x \in \mathcal{X}$ with the zero components of x .

Elaborating further on this, since $\langle g(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$ and any solution x^* of (VI), we directly infer that $-g(x^*)$ is an element of the normal cone $\text{NC}(x^*)$ to $\mathcal{X}$ at x^* . In our polyhedral setting, $\text{NC}(x^*)$ admits an especially simple representation as

$$\text{NC}(x^*) = \text{row}(A) - \{(v_1, \dots, v_n) \in \mathbb{R}_+^n : v_i = 0 \text{ whenever } x_i^* = 0\} \quad (7.61)$$

where $\text{row}(A) = (\ker A)^\perp \subseteq \mathbb{R}^n$ denotes the row space of A ([Hiriart-Urruty and Lemaréchal, 2001](#), Ex. 5.2.6). As a result, we see that x^* is a solution of (VI) if and only if $g(x^*)$ can be written in the form

$$g(x^*) - \sum_{i \in \mathcal{A}} v_i e_i \in \text{row}(A) \quad (7.62)$$

⁴ Inequality constraints of the form $Ax \leq b$ can also be accommodated in (7.60) by introducing the associated slack variables $s = b - Ax \geq 0$. Even though this leads to a more verbose presentation of $\mathcal{X}$, the form (7.60) is much more convenient in terms of notational overhead, so we stick with the equality formulation throughout.

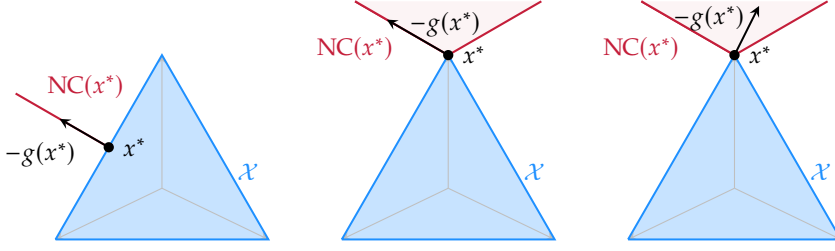


Figure 23: Different boundary solution configurations on the 2-dimensional unit simplex $\mathcal{X} = \{(x_1, x_2, x_3) \in \mathbb{R}_+^3 : x_1 + x_2 + x_3 = 1\}$ of $\mathbb{R}^3$: a non-extreme solution where g is sharp ($\mathcal{A} = \mathcal{A}_\# = \{1\}$, $\mathcal{A}_b = \emptyset$; left), an extreme solution where g is not sharp ($\mathcal{A} = \{1, 2\}$, $\mathcal{A}_\# = \{1\}$, $\mathcal{A}_b = \{2\}$; middle), and a sharp solution ($\mathcal{A} = \mathcal{A}_\# = \{1, 2\}$, $\mathcal{A}_b = \emptyset$; right).

for an ensemble of non-negative *slackness coefficients* $v_i \geq 0$, $i \in \mathcal{A}$, where

$$\mathcal{A} \equiv \mathcal{A}(x^*) = \{i : x_i^* = 0\} \quad (7.63)$$

denotes the set of inequality constraints of (7.60) that are active at x^* . With all this in mind, we distinguish the following solution configurations (see also Figure 23).

Definition 7.6 (Sharpness). Let $x^* \in \mathcal{X}$ be a solution of (VI) with associated slackness coefficients v_i , $i \in \mathcal{A}$, as per (7.62). The set of *sharp* ($\#$) and *flat* (b) directions at x^* are respectively defined as

$$\mathcal{A}_\# = \{i \in \mathcal{A} : v_i > 0\} \quad \text{and} \quad \mathcal{A}_b = \{i \in \mathcal{A} : v_i = 0\}, \quad (7.64)$$

and we say that g is *sharp* at x^* if $\mathcal{A}_\# = \mathcal{A}$ (or, equivalently, if $\mathcal{A}_b = \emptyset$). The *sharpness* of g at x^* is then defined as

$$v^* = \min_{i \in \mathcal{A}_\#} v_i \quad (7.65)$$

and, finally, if x^* is an extreme point of $\mathcal{X}$, we will say that x^* is itself *sharp*.

The terminology “sharp” and “flat” alludes to the case where g is a gradient field, and is best illustrated by an example. To wit, let $f(x_1, x_2) = x_1 + \frac{1}{2}(x_2 - 1)^2$ for $x_1, x_2 \geq 0$, so f admits a (unique) global minimizer at $x^* = (0, 1)$. Applying Definition 7.6 to $g = \nabla f$, we readily get $\mathcal{A}_\# = \{1\}$ and $\mathcal{A}_b = \{2\}$, reflecting the fact that $f(x_1, 1)$ exhibits a sharp minimum at 0 along x_1 whereas the landscape of $f(0, x_2)$ is flat to first-order around 1 along x_2 .

7.5.3 Convergence rate analysis

We are now in a position to state and prove our refinement of Theorem 7.2 for linearly constrained problems. To that end, following Alvarez et al. (2004), we will assume in this section that (BPM) is run with a Bregman regularizer h that is adapted to the polyhedral structure of $\mathcal{X}$ as per the definition below:

Definition 7.7. Let $\mathcal{X}$ be a polyhedral domain of the form (7.60). Then, a Bregman regularizer h on $\mathcal{X}$ is said to be *decomposable with kernel* θ if

$$h(x) = \sum_{i=1}^n \theta(x_i) \quad \text{for all } x \in \mathcal{X}. \quad (7.66)$$

for some continuous function $\theta: \mathbb{R}_+ \rightarrow \mathbb{R}$ such that (i) $\theta''(z) > 0$ for all $z > 0$; and (ii) θ'' is locally Lipschitz on $(0, \infty)$.

In addition to facilitating calculations, the notion of decomposability will further allow us to describe the convergence rate of the iterates of (BPM) near the boundary of $\mathcal{X}$ in finer detail. In fact, as it turns out, the speed of convergence along a given direction will actually be determined by the behavior of the derivative of the Bregman kernel θ near 0.

In this regard, there are two distinct regimes to consider. First, if $\lim_{x \rightarrow 0^+} \theta'(x) = -\infty$, it is straightforward to see that $\text{dom } \partial h = \text{ri } \mathcal{X}$ so, by Lemma E.2.1 in Appendix E, the iterates x_t of (BPM) will remain in $\text{ri } \mathcal{X}$ for all t ; in this case h is essentially smooth – or *Legendre* – in the sense of (Rockafellar, 1970, Chap. 26), and we will refer to it as *steep*. Otherwise, if $\theta'(0)$ exists and is finite, x_t may reach the boundary of $\mathcal{X}$ in a finite number of iterations; we will refer to this case as *non-steep*. The key difference between these two regimes is that, in the non-steep case, the algorithm may achieve convergence in a finite number of steps (at least along certain directions). On the other hand, even though finite-time convergence is not possible in the steep regime, the algorithm's rate of convergence may still depend on the boundary behavior of θ . To illustrate this, we will consider the following concrete cases:

Assumption 7.8. If $\theta: \mathbb{R}_+ \rightarrow \mathbb{R}$ is a kernel function as per Definition 7.7, θ' exhibits one of the following behaviors as $x \rightarrow 0^+$:

- (a) *Euclidean-like*: $\liminf_{x \rightarrow 0^+} \theta'(x) > -\infty$.
- (b) *Entropy-like*: $\liminf_{x \rightarrow 0^+} [\theta'(x) - \log x] > -\infty$.
- (c) *Power-like*: $\liminf_{x \rightarrow 0^+} x^p \theta'(x) > -\infty$ for some $p \in (0, 1)$.

Remark. Cases (a)–(c) above respectively mean that $|\theta'(x)|$ grows as $\mathcal{O}(1)$, $\mathcal{O}(|\log x|)$ or $\mathcal{O}(1/x^p)$ as $x \rightarrow 0^+$. Clearly, we have (a) $\implies$ (b) $\implies$ (c) so these cases are not exclusive; nonetheless, to avoid overloading the presentation, when we assume (b), we will tacitly imply that (a) does not also hold at the same time – and likewise for (c). $\heartsuit$

With all this in hand, we proceed below to show that, in linearly constrained problems, (BPM) converges along sharp directions at x^* at an accelerated rate relative to Theorem 7.2: sublinear rates may become linear, and linear rates transform to convergence in finite time.

Theorem 7.9. Suppose that (BPM) is run in a polyhedral domain with a decomposable regularizer as per Definition 7.7. Suppose further that Assumptions 7.1, 7.2, 7.4, and 7.8 hold, and that the method's step-size and initialization satisfy the requirements of Theorem 7.2. Then, for all $i \in \mathcal{A}_\sharp$, we have:

- (a) Under Assumption 7.8(a): there exists some $T \geq 1$ such that

$$x_{i,t} = 0 \quad \text{for all } t \geq T \quad (7.67a)$$

- (b) Under Assumption 7.8(b):

$$x_{i,t} = \mathcal{O}(\exp(-\gamma v_{\text{eff}} t / 2)) \quad (7.67b)$$

- (c) Under Assumption 7.8(c):

$$x_{i,t} = \mathcal{O}((\gamma v_{\text{eff}} t / 2)^{-1/p}) \quad (7.67c)$$

	Bregman Kernel (θ)	Bregman rate (Thm. 7.2)	Sharp rate (Thm. 7.9)
EUCLIDEAN	$x^2/2$	Linear	Finite time
ENTROPIC	$x \log x$	$\mathcal{O}(1/t)$	Linear
TSALLIS	$[q(1-q)]^{-1}(x - x^q)$	$\mathcal{O}(1/t^{q/(2-q)})$	$\mathcal{O}(1/t^{1/(1-q)})$
HELLINGER	$-\sqrt{1-x^2}$	$\mathcal{O}(1/t^{1/3})$	$\mathcal{O}(1/t^2)$

Table 2: Summary of the accelerated rates of convergence observed along sharp directions depending on the Bregman kernel (cf. Definition 7.7). The Euclidean, entropic and Tsallis kernels are the prototypical examples of Cases (a)–(c) of Assumption 7.8. The Bregman rate column refers to the Bregman divergence $D(x^*, x_t)$; the sharp rate refers to $x_{i,t}$ for a sharp direction $i \in \mathcal{A}_\sharp$ at x^* .

where

$$v_{\text{eff}} = \begin{cases} v^* & \text{if } g \text{ is sharp at } x^* \text{ (i.e., } \mathcal{A}_\flat = \emptyset), \\ v^*/\varrho & \text{otherwise,} \end{cases} \quad (7.68)$$

and $\varrho \equiv \varrho(A, b, x^*) \geq 1$ is a positive constant that depends only on $\mathcal{X}$ and x^* .

In particular, if x^* is sharp, we have the following immediate corollary of Theorem 7.9.

Corollary 7.10. *If x^* is sharp, then $\|x_t - x^*\| \rightarrow 0$ at a rate given by (7.67a), (7.67b), or (7.67c), depending respectively on whether Case (a), (b), or (c) of Assumption 7.8 holds.*

Proof. First, note that x^* is sharp if and only if $\text{span}\{e_i : i \in \mathcal{A}_\sharp\} + \text{row}(A) = \mathbb{R}^n$: indeed, since $\mathcal{X}$ is a polyhedron, x^* is extreme if and only if $\text{NC}(x^*)$ has nonempty topological interior, and this, combined with (7.61) and the fact that $\mathcal{A}_\sharp = \mathcal{A}$ (since g is sharp at x^*), proves our assertion. We thus conclude that, for all $i = 1, \dots, n$, there exist $\lambda_{ij} \in \mathbb{R}$, $j \in \mathcal{A}_\sharp$, such that $e_i - \sum_{j \in \mathcal{A}_\sharp} \lambda_{ij} e_j \in \text{row}(A)$, and hence, for all $t = 1, 2, \dots$, we have $x_{i,t} - x_i^* = \sum_{j \in \mathcal{A}_\sharp} \lambda_{ij} x_{j,t}$. Our claim then follows from Theorem 7.9 and the fact that all norms are equivalent on $\mathbb{R}^n$. ■

To facilitate comparisons with the non-sharp regime, the guarantees of Theorem 7.9 are juxtaposed with those of Theorem 7.2 in Table 2. Beyond this comparison, Theorem 7.9 is the main result of this section so, before proving it, some further remarks are in order.

Remark 1 (Solution configurations). By construction (and the fact that $\mathcal{X}$ admits a Slater point), it is straightforward to verify that g is sharp at x^* if and only if $g(x^*) \in -\text{ri}(\text{NC}(x^*))$; likewise, x^* is itself sharp if and only if $g(x^*) \in -\text{int}(\text{NC}(x^*))$. As we noted in the proof of Corollary 7.10, the latter condition is equivalent to asking that $\text{span}\{e_i : i \in \mathcal{A}_\sharp\} + \text{row}(A) = \mathbb{R}^n$, a condition which describes precisely the informal requirement that the sharp directions at x^* suffice to characterize it. By contrast, if g is sharp at some non-extreme point x^* , there exists some (nonzero) $z \in \text{TC}(x^*)$ such that $\langle g(x^*), z \rangle = 0$, indicating that the accelerated rates of Theorem 7.9 cannot be active along the residual direction z . We illustrate these distinct solution configurations in Figure 23. ☛

Remark 2 (Tightness and the structure of $\mathcal{X}$). We underline that the dependence of v_{eff} on the structure of $\mathcal{X}$ in the second branch of (7.68) cannot be lifted. To see this, let

$$\mathcal{X} = \{x \in \mathbb{R}_+^2 : x_1 = \varepsilon x_2\} \quad (7.69)$$

i.e., $A = [1 \ -\varepsilon]$ and $\ker A$ is spanned by the vector $z = (\varepsilon, 1)$. Then, if we take $g(x) = x - p$ with $p_1 \leq 0$ and $p_2 \geq 0$ (so that the origin is a solution),

and we equip $\mathcal{X}$ with the Bregman regularizer induced by the entropic kernel $\theta(x) = x \log x$, a straightforward calculation shows that the iterates of (MD) satisfy the recursion

$$\varepsilon \theta'(x_{1,t+1}) + \theta'(x_{2,t+1}) = [\varepsilon \theta'(x_{1,t}) + \theta'(x_{2,t})] + \gamma(\varepsilon x_{1,t} + x_{2,t}) - \gamma \langle p, z \rangle. \quad (7.70)$$

Thus, letting $\chi_t = x_{2,t} = x_{1,t}/\varepsilon$, the above can be rewritten as

$$\varepsilon \theta'(\varepsilon \chi_{t+1}) + \theta'(\chi_{t+1}) = [\varepsilon \theta'(\varepsilon \chi_t) + \theta'(\chi_t)] + \gamma(\varepsilon^2 + 1)\chi_t - \gamma \langle p, z \rangle \quad (7.71)$$

and hence, with $\theta'(x) = 1 + \log x$, we finally get

$$\chi_{t+1} = \chi_t \exp\left(-\gamma \frac{\varepsilon^2 + 1}{\varepsilon + 1} \chi_t + \gamma \frac{\langle p, z \rangle}{\varepsilon + 1}\right). \quad (7.72)$$

Now, if $\langle p, z \rangle < 0$, we readily infer that $\chi_t = x_{2,t}$ converges linearly to 0 at a rate of $\chi_t \sim \exp(-|\langle p, z \rangle|/(1 + \varepsilon) \cdot \gamma t)$ as predicted by Theorem 7.9. In particular, if $g(0) = -p = (v^*, v^*)$ with $v^* > 0$, the iterates of (MD) converge geometrically to zero with exponent γv^* , which matches the estimate of Theorem 7.9 up to a factor of 1/2 in the exponent. On the contrary, if $g(0) = -p = (v^*, 0)$, the term $|\langle -p, z \rangle| = \varepsilon v^*$ depends on the linear structure of $\mathcal{X}$ and can be arbitrarily bad as ε goes to zero. This illustrates why one cannot do away with the dependence on the linear structure of $\mathcal{X}$ when g is not sharp at x^* . $\blacksquare$

7.5.4 Proof of Theorem 7.9

We now proceed to the proof of Theorem 7.9, beginning with two helper lemmas tailored to the polyhedral structure of $\mathcal{X}$. The first is a book-keeping result regarding the subdifferentiability of h .

Lemma 7.11. *Let h be a decomposable regularizer on $\mathcal{X}$ with kernel θ as per Definition 7.7. Then the domain of subdifferentiability of h is $\mathcal{X}_h = \{x \in \mathcal{X} : x_i \in \text{dom } \partial \theta \text{ for all } i = 1, \dots, n\}$ and a continuous selection of ∂h is given by the expression*

$$\nabla h(x) = \sum_{i=1}^n \theta'(x_i) e_i \quad \text{for all } x \in \mathcal{X}_h. \quad (7.73)$$

Proof. See (Rockafellar, 1970, Thm 23.8), whose conditions are satisfied because $\mathcal{X}$ is polyhedral. $\blacksquare$

The second ingredient we will need is a separation result in the spirit of Farkas' lemma.

Lemma 7.12. *Let $\mathcal{X}$ be a polyhedral domain of the form (7.60). Then, for all $x^* \in \mathcal{X}$, there exists $P = P(A, b, x^*) \geq 1$ such that, for all $\mathcal{I} \subseteq \mathcal{A} \equiv \mathcal{A}(x^*)$, at least one of the following holds:*

- (a) $\mathcal{I} \neq \emptyset$ and there exists $i \in \mathcal{A} \setminus \mathcal{I}$ such that $x_i \leq P \max\{x_j : j \in \mathcal{I}\}$ for all $x \in \mathcal{X}$.
- (b) There exists $z \in \ker A$ such that $\|z\| \leq P$, $z_i = 0$ if $i \in \mathcal{I}$ and $P \geq z_i \geq 1$ if $i \in \mathcal{A} \setminus \mathcal{I}$.

The proof of Lemma 7.12 is based on Farkas' lemma so we relegate it to Appendix E. Armed with all this, we can finally proceed to prove our result for linearly constrained problems.

Proof of Theorem 7.9. We will consider two main cases, namely $\lim_{x \rightarrow 0^+} h'(x) = -\infty$ (the steep case) and $\lim_{x \rightarrow 0^+} h'(x) > -\infty$ (the non-steep case). The steep regime will cover Cases (b) and (c) of Assumption 7.8, whereas the non-steep regime will account for Case (a).

CASE 1: THE STEEP REGIME Note first that, without loss of generality, Cases (b) and (c) respectively imply that there exist $C \in \mathbb{R}$ and $\delta > 0$ such that, for all $x \in (0, \delta)$, we have:

$$\text{Under Assumption 7.8(b): } \theta'(x) \geq \log x - C \quad (7.74a)$$

$$\text{Under Assumption 7.8(c): } \theta'(x) \geq -Cx^{-p} \quad (7.74b)$$

With this in mind, let $r > 0$ be sufficiently small so that $\mathcal{B}_r := \{x \in \mathcal{X} : \|x - x^*\| \leq r\}$ satisfies:

- $\mathcal{B}_r \subseteq \mathcal{U} \cap \mathcal{B}$ with $\mathcal{B}$ and $\mathcal{U}$ defined by (SOS) and (7.23) respectively.
- If $x \in \mathcal{B}_r$ then $x_i < \delta$ for all $i \in \mathcal{A}$.
- If $x \in \mathcal{B}_r$, then $\|g(x) - g(x^*)\|_* \leq \nu^*/(2P)$ with ν^* by (7.65) and P by Lemma 7.12.

Now, recall that Step 1 of the proof of Theorem 7.2 implies that the iterates of (BPM) will remain in $\mathcal{B}_r$ for all t if x_1 is initialized sufficiently close to x^* . Given this stability guarantee, we will construct below two sets $\mathcal{I}_\# \subseteq \mathcal{A}_\#, \mathcal{I}_b \subseteq \mathcal{A}_b$ such that, for all $i \in \mathcal{I} := \mathcal{I}_\# \cup \mathcal{I}_b$, we have

$$x_{i,t} \leq P^{|\mathcal{I}|} \cdot \begin{cases} \exp(C + P\|\nabla h(x_1)\|_* + PR - \gamma v_{\text{eff}}(t-1)/2) & \text{under Assumption 7.8(b),} \\ C[\gamma v_{\text{eff}}(t-1)/2 - P\|\nabla h(x_1)\|_* - PR]_+^{-\frac{1}{p}} & \text{under Assumption 7.8(c),} \end{cases} \quad (7.75)$$

where $R := \sup_{x \in \mathcal{B}_r} \|\sum_{i \notin \mathcal{A}} \theta'(x_i) e_i\|_*$, $P \geq 1$ is the constant given by Lemma 7.12 and v_{eff} is defined as in (7.68) with $\varrho = P|\mathcal{A}|$.

Our construction proceeds inductively, starting with $\mathcal{I}_\# = \mathcal{I}_b = \emptyset$, for which the stated property holds trivially. For the inductive step, if (7.75) holds for $\mathcal{I}_\# \subseteq \mathcal{A}_\#$ and $\mathcal{I}_b \subseteq \mathcal{A}_b$, we will show that there exists some $j \in \mathcal{A} \setminus \mathcal{I}$ such that (7.75) still holds for $\mathcal{I} \cup \{j\}$. Since the number of active constraints is finite, the progressive addition of these indices will allow us to reach $\mathcal{I}_\# = \mathcal{A}_\#$, thus proving our initial claim.

To carry all this out, assume that $\mathcal{I}_\# \subseteq \mathcal{A}_\#, \mathcal{I}_b \subseteq \mathcal{A}_b$, and apply Lemma 7.12 to $\mathcal{I} = \mathcal{I}_\# \cup \mathcal{I}_b$. If the first case of Lemma 7.12 holds, then $\mathcal{I} \neq \emptyset$ and there exists $i \in \mathcal{A}$ such that

$$x_{i,t} \leq P \max_{j \in \mathcal{I}} x_{j,t} \quad (7.76)$$

so (7.75) still holds when i is appended to $\mathcal{I}_\#$ or $\mathcal{I}_b$. Otherwise, the second case of Lemma 7.12 holds and there exists $z \in \ker A$ with $\|z\| \leq P$, $z_i = 0$ if $i \in \mathcal{I}$, and $P \geq z_i \geq 1$ if $i \in \mathcal{A} \setminus \mathcal{I}$. Since h is steep, this means that $x_{t+1} = P_{x_t}(-\gamma g_{t+1/2})$ belongs to $\mathcal{X}_h = \text{ri } \mathcal{X}$ for all $t = 1, 2, \dots$, so $\text{NC}(x_{t+1})$ is the affine hull of $\mathcal{X}$, i.e., $\text{NC}(x_{t+1}) = \text{row}(A)$. Hence, Lemma E.2.1 in Appendix E guarantees

$$\nabla h(x_{t+1}) - \nabla h(x_t) + \gamma g_{t+1/2} \in \text{row}(A) \quad (7.77)$$

so, telescoping from $s = 1$ to $t - 1$, we get $\nabla h(x_t) - \nabla h(x_1) + \gamma \sum_{s=1}^{t-1} g_{s+1/2} \in \text{row}(A)$. Thus, taking the scalar product with $z \in \ker A$ yields

$$\sum_{i=1}^n \theta'(x_{i,t}) z_i = \langle \nabla h(x_1), z \rangle - \gamma \sum_{s=1}^{t-1} \langle g_{s+1/2}, z \rangle \quad (7.78)$$

so, after rearranging and invoking (7.62) to write $\langle g(x^*), z \rangle = \sum_{i \in \mathcal{A}} v_i z_i$, we get

$$\begin{aligned} \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \theta'(x_{i,t}) z_i + \sum_{i \in \mathcal{I}} \theta'(x_{i,t}) z_i &= \langle \nabla h(x_1), z \rangle - \gamma \sum_{s=1}^{t-1} \sum_{i \in \mathcal{A}} v_i z_i \\ &\quad + \gamma \sum_{s=1}^{t-1} \langle g(x^*) - g_{s+1/2}, z \rangle - \sum_{i \notin \mathcal{A}} \theta'(x_{i,t}) z_i. \end{aligned} \quad (7.79)$$

Finally, by the properties we used to construct z , we further have

$$\begin{aligned} \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \theta'(x_{i,t}) z_i &\leq P \|\nabla h(x_1)\|_* - \gamma(t-1) \sum_{i \in \mathcal{A} \setminus \mathcal{I}} v_i z_i \\ &\quad + \gamma \sum_{s=1}^{t-1} P \|g(x^*) - g_{s+1/2}\|_* + P \left\| \sum_{i \notin \mathcal{A}} \theta'(x_{i,t}) e_i \right\|_* \\ &\leq P \|\nabla h(x_1)\|_* - \gamma(t-1) \sum_{i \in \mathcal{A} \setminus \mathcal{I}} v_i z_i + \frac{1}{2} \gamma v^*(t-1) + PR. \end{aligned} \quad (7.80)$$

where the second inequality follows from how we chose $\mathcal{B}_r$ at the beginning of the proof.

We conclude by distinguishing whether g is sharp at x^* (i.e., if $\mathcal{A}_b = \emptyset$ or not).

Case 1: If $\mathcal{A}_b = \emptyset$, we have $\mathcal{A} \setminus \mathcal{I} = \mathcal{A}_\# \setminus \mathcal{I}_\#$ and $v_i \geq v^*$ for all $i \in \mathcal{A} \setminus \mathcal{I}$, so (7.80) gives

$$\sum_{i \in \mathcal{A} \setminus \mathcal{I}} \theta'(x_{i,t}) z_i + \gamma(t-1) v^* z_i \leq P \|\nabla h(x_1)\|_* + \frac{1}{2} \gamma v^*(t-1) + PR. \quad (7.81)$$

Choosing the coordinate $j \in \mathcal{A} \setminus \mathcal{I}$ corresponding to the smallest term in the sum on the left-hand side (LHS), we get

$$(\theta'(x_{j,t}) + \gamma(t-1)v^*) (| \mathcal{A} \setminus \mathcal{I} | z_j) \leq P \|\nabla h(x_1)\|_* + \frac{1}{2} \gamma v^*(t-1) + PR \quad (7.82)$$

and noting that $| \mathcal{A} \setminus \mathcal{I} | z_j \geq 1$ yields

$$\theta'(x_{j,t}) \leq P \|\nabla h(x_1)\|_* - \frac{1}{2} \gamma v^*(t-1) + PR. \quad (7.83)$$

Case 2: If $\mathcal{A}_b \neq \emptyset$, then, since $\mathcal{I}_\# \subsetneq \mathcal{A}_\#$, the intersection of $\mathcal{A} \setminus \mathcal{I}$ and $\mathcal{A}_\#$ is not empty so that, $\sum_{i \in \mathcal{A} \setminus \mathcal{I}} v_i z_i \geq \sum_{i \in \mathcal{A} \setminus \mathcal{I}} v_i \geq v^*$ and the inequality (7.80) above becomes,

$$\sum_{i \in \mathcal{A} \setminus \mathcal{I}} \theta'(x_{i,t}) z_i \leq P \|\nabla h(x_1)\|_* - \frac{1}{2} \gamma v^*(t-1) + PR. \quad (7.84)$$

Now, choosing j to be the coordinate in $\mathcal{A} \setminus \mathcal{I}$ which minimizes the LHS and bounding $|\mathcal{A} \setminus \mathcal{I}|$ by 1 and $|\mathcal{A}|$, we get that

$$\theta'(x_{j,t})z_j \leq P\|\nabla h(x_1)\|_* - \frac{\gamma v^*}{2|\mathcal{A}|}(t-1) + PR. \quad (7.85)$$

Dividing both sides by z_j and using that it lies between 1 and P gives

$$\theta'(x_{j,t}) \leq P\|\nabla h(x_1)\|_* - \frac{\gamma v^*}{2P|\mathcal{A}|}(t-1) + PR \quad (7.86)$$

Thus, in both cases, there exists $j \in \mathcal{A} \setminus \mathcal{I}$ such that (7.86) holds, since $P|\mathcal{A}| \geq 1$. Therefore, combining this inequality with (7.74) we conclude that (7.75) holds for j , and since $P \geq 1$, we conclude that we can augment $\mathcal{I}_\#$ or $\mathcal{I}_b$ by j , depending on whether it belongs to $\mathcal{A}_\#$ or $\mathcal{A}_b$.

CASE 2: THE NON-STEEP REGIME The proof borrows the structure of the first case, though it is more direct. Take $r > 0$ small enough such that $\mathcal{B}_r := \{x \in \mathcal{X} : \|x - x^*\| \leq r\}$ satisfies

- $\mathcal{B}_r$ is included in $\mathcal{U} \cap \mathcal{B}$, and if $x \in \mathcal{B}_r$, then $\|g(x) - g(x^*)\|_* \leq \frac{v^*}{3P}$,
- if $x, x' \in \mathcal{B}_r$ then, $\|\nabla h(x) - \nabla h(x')\|_* \leq \frac{\gamma v^*}{3P}$ where P is given by Lemma 7.12. This is possible since ∇h is continuous at x^* ,
- No other constraint $x_i = 0$ with $i \notin \mathcal{A}$ becomes active in $\mathcal{B}_r$.

As we have seen in the proof of Theorem 7.2, if (BPM) is initialized close enough to x^* , then all the iterates x_t and the half-iterates $x_{t+1/2}$ for $s = 1, 2, \dots$ are contained in $\mathcal{B}_r$.

As above, fix some $t \geq T$. We will build sets $\mathcal{I}_\# \subseteq \mathcal{A}_\#, \mathcal{I}_b \subseteq \mathcal{A}_b$ with the property that that

$$\text{for all } i \in \mathcal{I}_\# \cup \mathcal{I}_b, x_{i,t} = 0. \quad (7.87)$$

Starting with $\mathcal{I}_\# = \mathcal{I}_b = \emptyset$, (7.87) is trivially verified. Now, take $\mathcal{I}_\# \subseteq \mathcal{A}_\#, \mathcal{I}_b \subseteq \mathcal{A}_b$ which satisfy the desired property and, as before, apply Lemma 7.12 with $\mathcal{I} := \mathcal{I}_\# \cup \mathcal{I}_b$. If the first case of Lemma 7.12 holds, then $\mathcal{I} \neq \emptyset$ and there exists $j \in \mathcal{A}$ such that,

$$x_{j,t} \leq P \max(x_{i,t} : i \in \mathcal{I}) \quad (7.88)$$

which yields the result by adding i to $\mathcal{I}_\#$ or $\mathcal{I}_b$ depending whether it belongs to $\mathcal{A}_\#$ or $\mathcal{A}_b$. Otherwise, if the second case holds, there is some $z \in \ker A$ such that $\|z\| \leq P, z_i = 0$ if $i \in \mathcal{I}$ and $P \geq z_i \geq 1$ if $i \in \mathcal{A} \setminus \mathcal{I}$. For the sake of contradiction, assume that for all $i \in \mathcal{A} \setminus \mathcal{I}, x_{i,t} > 0$. Showing that this results in a contradiction will give us an additional coordinate $j \in \mathcal{A} \setminus \mathcal{I}$ for which $x_{j,t} = 0$ that we will then add to $\mathcal{I}_\#$ or $\mathcal{I}_b$ as in the first case.

Now, let us determine the normal cone at x_t . Since x_t belongs to $\mathcal{B}_r^{\mathcal{X}}(x^*)$, no other constraint other than the ones of $\mathcal{I}$ can become active, and these constraints are actually active by the definition of $\mathcal{I}$ and (7.87). Hence, the normal cone at x_t (see (7.61)) becomes

$$\text{NC}(x_t) = \left\{ - \sum_{i \in \mathcal{I}} v_i e_i : (v_i)_{i \in \mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{I}} \right\} + \text{row}(A). \quad (7.89)$$

Taking a scalar product between the last inclusion of [Lemma E.2.1](#) in [Appendix E](#) and z , we get that

$$\langle \nabla h(x_t) - \nabla h(x_{t-1}) - \gamma g_{t-1/2}, z \rangle = 0. \quad (7.90)$$

This means, from the definition of $\mathcal{B}_r^{\mathcal{X}}(x^*)$, that,

$$\gamma \langle g(x^*), z \rangle = \langle \nabla h(x_t) - \nabla h(x_{t-1}), z \rangle + \gamma \langle g(x^*) - \gamma g_{t-1/2}, z \rangle \leq \frac{2\gamma v^*}{3} \quad (7.91)$$

However, by [\(7.62\)](#) and the properties of z , we also have

$$\langle g(x^*), z \rangle = \sum_{i \in \mathcal{A} \setminus \mathcal{I}} v_i z_i \geq v^* \quad (7.92)$$

which is in contradiction with [\(7.91\)](#). We may therefore iteratively add coordinates of $\mathcal{A}$ for which $x_{i,t} = 0$, which completes the induction and thus our proof. $\blacksquare$

7.6 MIRROR DESCENT ON STOCHASTIC VARIATIONAL INEQUALITIES

7.6.1 The Stochastic optimistic mirror descent algorithm

In this section, we recall the *optimistic mirror descent* (OMD) method for solving [\(VI\)](#).

The oracle model

We focus on *stochastic first-order oracles* of the form

$$\text{Or}(x; \theta) = g(x) + \text{Err}(x; \theta) \quad (\text{SFO})$$

where θ is a random variable taking values in some abstract measure space Θ and $\text{Err}(x; \theta)$ is an umbrella error term capturing all sources of uncertainty in the model.

In practice, the oracle is called repeatedly at a sequence of query points $x_t \in \mathcal{X}$ with a different random seed θ_t at each time.⁵ The sequence of oracle signals $g_t = \text{Or}(x_t; \theta_t)$ may then be written as

$$g_t = g(x_t) + U_t \quad (7.93)$$

where $U_t = \text{Err}(x_t; \theta_t)$ denotes the error of the oracle model at stage t . To keep track of this sequence of events, we will treat x_t as a stochastic process on some complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and we will write $\mathcal{F}_t := \mathcal{F}(x_1, \dots, x_t) \subseteq \mathcal{F}$ for the *history of play* up to stage t (inclusive). Since the randomness entering the oracle is triggered only *after* a query point has been selected, we will posit throughout that θ_t (and hence U_t and g_t) is $\mathcal{F}_{t+1}$ -measurable – but not necessarily $\mathcal{F}_t$ -measurable. We will also make the blanket assumption below for the oracle.

Assumption 7.9 (Oracle signal). The error term U_t in [\(7.93\)](#) satisfies the following properties:

$$a) \quad \text{Zero-mean:} \quad \mathbb{E}[U_t \mid \mathcal{F}_t] = 0. \quad (7.94a)$$

⁵ In the sequel, we will allow the index t to take both integer and half-integer values.

$$b) \text{ Finite variance: } \mathbb{E}[\|U_t\|_*^2 \mid \mathcal{F}_t] \leq \sigma^2. \quad (7.94b)$$

Both assumptions are standard in the literature on stochastic methods, cf. [Polyak \(1987\)](#), [Hazan \(2012\)](#), [Bubeck \(2015\)](#), and references therein.

With all this in hand, the optimistic mirror descent algorithm is defined in recursive form as follows:

$$\begin{aligned} x_{t+1/2} &= P_{x_t}(-\gamma_t g_{t-1/2}) \\ x_{t+1} &= P_{x_t}(-\gamma_t g_{t+1/2}) \end{aligned} \quad (\text{OMD})$$

In the above, $t = 1, 2, \dots$, is the algorithm's iteration counter, x_t and $x_{t+1/2}$ respectively denote the algorithm's *base* and *leading* states at stage t , $g_{t+1/2}$ is an oracle signal obtained by querying (SFO) at $x_{t+1/2}$, and $P_x(y)$ denotes the method's so-called "prox-mapping". In terms of initialization, we also take $x_1 = x_{1/2} \in \text{ri } \mathcal{X}$ for simplicity. All these elements are defined in detail below.

At a high level, (OMD) seeks to leverage past feedback to anticipate the landscape of g and perform more informed steps in subsequent iterations. In more detail, starting at some base state x_t , $t = 1, 2, \dots$, the algorithm first generates an intermediate, leading state $x_{t+1/2}$, and then updates the base state with oracle input from $x_{t+1/2}$ – that is, $g_{t+1/2}$. This is also the main idea behind the extra-gradient algorithm of [Korpelevich \(1976\)](#); the key difference here is that OMD avoids making two oracle queries per iteration by using the oracle input $g_{t-1/2}$ received at the previous leading state $x_{t-1/2}$ as a proxy for g_t (which is never requested or received by the optimizer).

7.6.2 Analysis and results

We are finally in a position to state and prove our main result for the convergence rate of (OMD). Since we make no global monotonicity assumptions for the defining vector field g , this result is *de facto* of a local nature. We provide below the complete statement, followed by a series of remarks and an outline of the main steps of the proof.

Theorem 7.13. *Let x^* be a solution of (VI) and fix some confidence level $\delta > 0$. Suppose further that [Assumptions 7.1, 7.2, and 7.9](#) hold and (OMD) is run with a step-size of the form $\gamma_t = \gamma/(t + t_0)^\eta$ with $\eta \in (1/2, 1]$ and $\gamma, t_0 > 0$. Then, (OMD) enjoys the following properties:*

STOCHASTIC STABILITY. *If γ/t_0 is small enough, x^* is stochastically stable: for every neighborhood $\mathcal{U}$ of x^* , there exists a neighborhood $\mathcal{U}_1$ of x^* such that the event*

$$E_{\mathcal{U}} = \{x_t \in \mathcal{U} \text{ for all } t = 1, 2, \dots\} \quad (7.95)$$

occurs with probability at least $1 - \delta$ if $x_1 \in \mathcal{U}_1$.

CONVERGENCE. *x_t converges to x^* at the following rate:*

Case I: *If the Legendre exponent of h at x^* is $\beta \in (0, 1)$ and $\min\{\gamma, 1/t_0\}$ is small enough, the iterates x_t of (OMD) enjoy the convergence rate*

$$\mathbb{E}[D(x^*, x_t) \mid E_{\mathcal{U}}] = \begin{cases} \mathcal{O}^*(\log t)^{-(1-\beta)/\beta} & \text{if } \eta = 1, \\ \mathcal{O}^* t^{-\min\{(1-\eta)(1-\beta)/\beta, \eta\}} & \text{if } \frac{1}{2} < \eta < 1, \end{cases} \quad (7.96)$$

provided that $\mathcal{U}$ is small enough and $x_1 \in \mathcal{U}_1$. In particular, this rate undergoes a phase transition at $\beta = 1/2$:

- *Euclidean-like phase*: If $\beta \in (0, 1/2)$ and $\eta = 1 - \beta$, we obtain the optimized rate

$$\mathbb{E}[D(x^*, x_t) | E_{\mathcal{U}}] = \mathcal{O} * t^{-(1-\beta)}. \quad (7.97)$$

- *Entropy-like phase*: If $\beta \in [1/2, 1)$ and $\eta = (1 + \epsilon)/2$ for any small $\epsilon > 0$, we get

$$\mathbb{E}[D(x^*, x_t) | E_{\mathcal{U}}] = \mathcal{O} * t^{-\frac{1-\epsilon}{2\beta}(1-\beta)}. \quad (7.98)$$

Case II: If $\beta = 0$ and γ and t_0 are sufficiently large, the iterates x_t of (OMD) enjoy the rate

$$\mathbb{E}[D(x^*, x_t) | E_{\mathcal{U}}] = \mathcal{O} t^{-\eta}, \quad (7.99)$$

provided that $\mathcal{U}$ is small enough and $x_1 \in \mathcal{U}_1$. In particular, for $\eta = 1$ we get

$$\mathbb{E}[D(x^*, x_t) | E_{\mathcal{U}}] = \mathcal{O} t^{-1}. \quad (7.100)$$

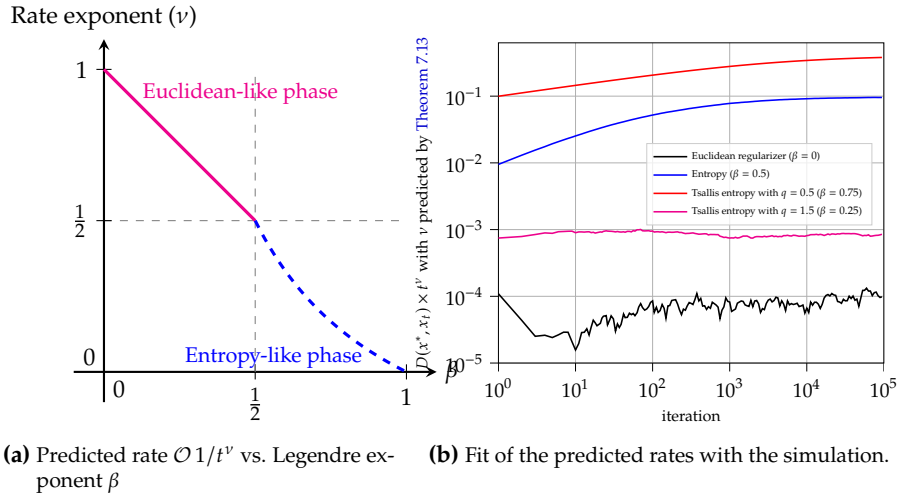


Figure 24: Illustration of the rates predicted by Theorem 7.13 (left: theory; right: numerical validation). The dashed line in Figure 24a indicates that the rate prediction is only valid up to an arbitrarily small exponent or suitable logarithmic factor. The experiments in Figure 24b were conducted for a linear vector field $g(x) = x$ on $\mathcal{X} = [0, +\infty)$ with the regularizer setups of Examples 7.1–7.3 (for more details, see F.6). We run OMD with the stepsizes prescribed by Theorem 7.13 and plot $D(x^*, x_t) \times t^\nu$ where ν is the predicted rate ($\nu = 1, 0.5, 0.167, 0.75$ respectively for Euclidean, entropy, Tsallis entropy with $q = 0.5$ and $q = 1.5$). A horizontal line corresponds to a match between the theoretical and observed rates.

RELATED WORK The rate guarantees of Theorem 7.13 should be compared and contrasted to a series of recent results on the convergence speed of OMD and its variants. Most of these results concern the deterministic case, i.e., when $U_t = 0$ for all t . In this regime, and focusing on the last iterate of the method, Golowich et al. (2020a,b) showed that the convergence speed of the (optimistic) extra-gradient algorithm in unconstrained, smooth, monotone variational inequalities is $\mathcal{O} 1/\sqrt{t}$ in terms of $\|g(x_t)\|$. If, in addition, g is strongly monotone, it is well known that this rate becomes linear; for a recent take on different (Euclidean) variants of (OMD), see Malitsky (2015), Gidel et al. (2019), Hsieh et al. (2019), Mokhtari et al. (2019a,b), and references therein.

Unsurprisingly, the stochastic regime is fundamentally different. In the merely monotone case (i.e., neither strongly nor strictly monotone), we are not aware of any result on the method’s last-iterate convergence speed: existing results in the literature either cover the algorithm’s ergodic average (Juditsky et al., 2011, Gidel et al., 2019, Cui and Shanbhag, 2016) or are asymptotic in nature (Iusem et al., 2017). Because of the lack of deterministic correlation between gradients at the base and leading states in stochastic problems, smoothness does not help to achieve better rates in this case – except for mollifying the requirement for bounded gradient signals (Juditsky et al., 2011). By contrast, strong monotonicity – local or global – *does* help: as was shown by Hsieh et al. (2019), the iterates of the *Euclidean* version of (OMD) run with a $1/t$ step-size schedule converge locally to solutions satisfying (SOS) at a $\mathcal{O} 1/t$ rate in terms of the mean square distance to the solution. This result is a special instance of Case II of Theorem 7.13: in particular, for $\beta = 0, \eta = 1$, (7.99) indicates that the $1/t$ step-size schedule examined by Hsieh et al. (2019) is optimal for the Euclidean case; however, as we explain below, it is *not* optimal for non-Euclidean distance-generating functions (DGFs).

STEP-SIZE TUNING A key take-away from Theorem 7.13 is that the step-size schedule that optimizes the rate guarantees in (7.97) and (7.98) is either $\gamma_t = \Theta(1/t^{1-\beta})$ when $0 \leq \beta < 1/2$, or as close as possible to $\Theta(t^{1/2})$ when $1/2 \leq \beta < 1$; in both cases, the optimal tuning of the algorithm’s step-size *depends on the Legendre exponent of x^** . We feel that this is an important parameter to keep in mind when deploying non-Euclidean versions of (OMD) in a stochastic setting – e.g., as in the optimistic multiplicative weights update (OMWU) algorithm of Daskalakis and Panageas (2019). Theorem 7.13 suggests that the “best practices” of the Euclidean setting do not carry over to non-Euclidean ones: Euclidean step-size policy $1/t$ would lead to a catastrophic drop from $1/\sqrt{t}$ to $1/\log t$ in the rate guarantees for (OMD).

GEOMETRY-DEPENDENT RATES One final observation for Theorem 7.13 is that the convergence speed of (OMD) is measured in terms of the Bregman divergence $D(x^*, x_t)$. Since $\|x_t - x^*\| = \mathcal{O} \sqrt{D(x^*, x_t)}$, this rate immediately translates to a rate on $\|x_t - x^*\|$. However, this rate is *not* tight, again because of the disparity between the Bregman and norm geometries discussed in § 7.3. Specifically, in many cases, a non-zero Legendre exponent implies a sharper lower bound on the Bregman divergence of the form $D(x^*, x_t) = \Omega(\|x_t - x^*\|^{2-2\beta})$. Thus, for $\beta \in [0, 1/2)$, Jensen’s inequality applied to the optimized bound (7.97) gives $\mathbb{E}[\|x_t - x^*\|] = \mathcal{O} 1/\sqrt{t}$. Importantly, this rate guarantee is the same as in the Euclidean case ($\beta = 0$), though it requires a completely different step-size policy to achieve it. Determining when the Legendre exponent also provides a lower bound is a very interesting direction for future research, but one that lies beyond the scope of this work.

7.6.3 Main ideas of the proof

The heavy lifting in the proof of Theorem 7.13 is provided by Proposition 7.14 below, which also makes explicit some of the hidden constants in the statement of the theorem.

Proposition 7.14. *With notation and assumptions as in [Theorem 7.13](#), fix some $r > 0$ and let*

$$\mathcal{U} = \mathcal{B}(x^*, r) \cap \mathcal{X} \quad \text{and} \quad \mathcal{U}_1 = \{x : D(x^*, x) \leq r^2/12\}.$$

If $x_1 \in \mathcal{U}_1$, then the event $E_{\mathcal{U}}$ defined in [\(7.95\)](#) satisfies $\mathbb{P}(E_{\mathcal{U}}) \geq 1 - \mathcal{O}^*\rho^2/r^2$ where $\rho^2 = \sigma^2\gamma^2t_0^{1-2\eta}/(2\eta - 1)$. Moreover, if r is taken small enough, $\mathbb{E}[D(x^*, x_t) \mid E_{\mathcal{U}}]$ is bounded according to the following table and conditions:

Legendre exponent	Rate ($\eta = 1$)	Rate ($\frac{1}{2} < \eta < 1$)
$\beta \in (0, 1)$	$\mathcal{O}^*(\log t)^{-\frac{1-\beta}{\beta}}$	$\mathcal{O}^*t^{-\frac{(1-\eta)(1-\beta)}{\beta}} + t^{-\eta}$
Conditions:	$\gamma^{1+1/\beta} \leq \frac{c(\eta, \alpha) \max(1, \Phi)}{4\sigma^2(\mu/K)^{1/\beta-1}}$	
$\beta = 0$	$\mathcal{O} 1/t$	$\mathcal{O} 1/t^\eta$
Conditions:	$\gamma > K/\mu$	—

where $\alpha = \beta/(1 - \beta)$, $\Phi = r^2/12 + 8\rho^2$, and

$$c(\eta, \alpha) = \begin{cases} \frac{1-\eta}{(1+\alpha)^{1+1/\alpha}} & \text{if } \eta \leq \frac{1+\alpha}{1+2\alpha}, \\ \alpha \left(\frac{1-2^{-(1+\alpha)}}{1+\alpha} \right)^{1+1/\alpha} & \text{if } \eta > \frac{1+\alpha}{1+2\alpha}. \end{cases} \quad (7.101)$$

We outline here the main ideas of the proof, deferring all details to [Appendix F.5](#).

Sketch of proof. Our proof strategy is to show that, under the stated conditions for r and the algorithm's initialization, the iterates of (OMD) remain within a neighborhood of x^* where (SOS) holds with high probability. Then, conditioning on this event, we use the Bregman divergence as a stochastic potential function, and we derive the algorithm's rate of convergence using a series of lemmas on (random) numerical sequences. The step-by-step process is as follows:

1. The cornerstone of the proof is a descent inequality for the iterates of OMD, which relates $D(x^*, x_{t+1})$ to $D(x^*, x_t)$. As stated in [Lemma F.3.2](#), we have

$$D(x^*, x_{t+1}) + \phi_{t+1} \leq D(x^*, x_t) + (1 - \gamma_t\mu)\phi_t - \gamma_t \langle g_{t+1/2}, x_{t+1/2} - x^* \rangle \quad (7.102)$$

$$+ \left(4\gamma_t^2 L^2 - \frac{1}{2} \right) \|x_{t+1/2} - x_t\|^2 + 4\gamma_t^2 [\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2]$$

where $\phi_t = \frac{\gamma_{t-1}^2}{2} \|g_{t-1/2} - g_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = 0$.

2. The next step is to show that, with high probability and if initialized close to x^* , the iterates of OMD remain in $\mathcal{U}$. This is a technical argument relying on a use of the Doob-Kolmogorov maximal inequality for submartingales (in the spirit of [Hsieh et al. \(2020\)](#)) and the condition $\eta > 1/2$ (which in turn guarantees that the step-size is square summable, so the corresponding submartingale error terms are bounded in L^1). The formal statement is [Lemma F.4.2](#).
3. To proceed, choose $r > 0$ small enough so that the second order sufficiency condition (SOS) holds and the Legendre exponent estimate [\(7.19\)](#) both hold. More precisely, we take r small enough so that $\mathcal{U} = \mathcal{B}(x^*, r) \cap \mathcal{X}$ is included in the domains of validity of both properties ($\mathcal{B}$ and $\mathcal{U}$ respectively), and we will

work on the event $E_t = \{x_{s+1/2} \in \mathcal{U} \text{ for all } s = 1, 2, \dots, t\}$. Then, conditioning on E_t , [Assumption 7.2](#) gives

$$\langle g(x_{t+1/2}), x_{t+1/2} - x^* \rangle \geq \frac{\mu}{2} \|x_{t+1/2} - x^*\|^2,$$

and hence, by [Lemma F.1.6](#), we get

$$\langle g(x_{t+1/2}), x_{t+1/2} - x^* \rangle \geq \frac{\mu}{2} \|x_t - x^*\|^2 - \mu \|x_{t+1/2} - x_t\|^2.$$

However, the descent inequality only involves $D(x^*, x_t)$ and not $\|x_t - x^*\|^2$, so we need to relate the latter to the former. This is where the Legendre exponent comes into play. More precisely, [Definition 7.5](#) with $1 + \alpha = \frac{1}{1-\beta}$ gives

$$D(x^*, x)^{1+\alpha} \leq \frac{1}{2} K \|x^* - x\|^2,$$

so, with $x_{t+1/2} \in \mathcal{U}$, we have

$$\langle g(x_{t+1/2}), x_{t+1/2} - x^* \rangle \geq \frac{\mu}{K} D(x^*, x_t)^{1+\alpha} - \mu \|x_{t+1/2} - x_t\|^2.$$

4. Thus, going back to the descent inequality (7.102), employing the bound derived above, and taking expectations ultimately yields (after some technical calculations) the iterative bound

$$\mathbb{E} \left[(D(x^*, x_{t+1}) + \phi_{t+1}) \mathbb{1}_{E_t} \right] \quad (7.103)$$

$$\leq \mathbb{E} \left[\left(D(x^*, x_t) - \frac{\mu \gamma_t}{K} D(x^*, x_t)^{1+\alpha} + \left(1 - \frac{\gamma_t \mu}{K} \right) \phi_t \right) \mathbb{1}_{E_{t-1}} \right] + 8\gamma_t^2 \sigma^2 \quad (7.104)$$

where we used [Assumption 7.9](#) to bound the expectations of $\|U_{t+1/2}\|_*^2$ and $\|U_{t-1/2}\|_*^2$.

5. We are now in a position to reduce the problem under study to the behavior of the sequence $a_t = \mathbb{E} \left[(D(x^*, x_t) + \phi_t) \mathbb{1}_{E_{t-1}} \right]$, $t \geq 1$. Indeed, taking a closer look at E_t shows that

$$\mathbb{E}[D(x^*, x_t) \mid E_{\mathcal{U}}] = \mathcal{O} a_t.$$

Consequently, the inequality (7.104) above can be rewritten as follows:

$$a_{t+1} \leq a_t - \frac{\mu \gamma_t}{K} \mathbb{E} \left[\left(D(x^*, x_t)^{1+\alpha} + \phi_t \right) \mathbb{1}_{E_{t-1}} \right] + 8\gamma_t^2 \sigma^2. \quad (7.105)$$

6. The behavior of this last sequence hinges heavily on whether $\alpha > 0$ or not. In detail, we have:

- If $\alpha = 0$, (7.105) gives

$$a_{t+1} \leq \left(1 - \frac{\mu \gamma}{K(t+t_0)^\eta} \right) a_t + \frac{8\gamma^2 \sigma^2}{(t+t_0)^{2\eta}}.$$

The long-run behavior of this recursive inequality is described by [Lemmas F.2.1](#) and [F.2.2](#): as long as $\mu \gamma / \beta > 1$ for $\eta = 1$ (and only for $\eta = 1$), we have $a_t = \mathcal{O} 1/t^\eta$.

- If $\alpha > 0$, a series of further technical calculations in the same spirit yields

$$a_{t+1} \leq a_t - \frac{\mu \gamma}{2^\alpha \max(1, \Phi^\alpha) K(t+t_0)^\eta} a_t^{1+\alpha} + \frac{8\gamma^2 \sigma^2}{(t+t_0)^{2\eta}}$$

The final step of our proof is to upper bound the behavior of a_t based on this recursive inequality. The necessary groundwork is provided by [Lemmas F.2.4](#) and [F.2.5](#), and hinges on whether $\eta \geq \frac{1+\alpha}{1+2\alpha}$ or not (and is also where the value of $c(\eta, \alpha)$ comes in).

Putting everything together, we obtain the conditional rates in the statement of the proposition. ■

7.7 CONCLUSION

Our results indicate that Euclidean regularization leads to faster last-iterate convergence rates near second-order sufficient (SOS) solutions. While this does not contradict the analysis of [Nemirovski \(2004\)](#) – which concerns the method’s ergodic average and advocates the use of non-Euclidean regularizers in domains with a favorable geometry – it *does* run contrary to its spirit. We attribute this discrepancy to the fact that Lipschitz continuity and second-order sufficiency are both norm-based conditions, so it is plausible to expect that norm-based regularizers lead to better last-iterate behavior.

We further investigated this relationship in the stochastic setting, extending the analysis to optimistic mirror descent with stochastic oracle feedback. There too, the Legendre exponent of the distance-generating function governs the last-iterate convergence rate: the less regular the Bregman divergence around the solution, the slower the convergence of the last iterate. In particular, the optimal step-size policy depends on the distance-generating function through the Legendre exponent.

These results open several directions for future work. A natural question is what the corresponding rate analysis gives for Bregman-based variants of the Lipschitz and strong monotone conditions, e.g., as in [Bauschke et al. \(2017\)](#), [Lu et al. \(2018\)](#), [Antonakopoulos et al. \(2019\)](#).

PERSPECTIVES

There are several open questions and directions for future work that open up as a result of the material presented in this manuscript. These can be organized around two overarching objectives that are the primary driving forces behind my work:

- (i) Gaining further insight into the training dynamics of deep learning models;
- (ii) Designing machine learning systems that we can trust, from optimization to uncertainty quantification.

I believe that, over time, these two research directions will be tightly intertwined: having a theoretical understanding of the training dynamics of deep learning models will allow us to design grounded robust training and inference procedures. But for now, I will describe them separately.

8.1 TOWARDS A PRINCIPLED UNDERSTANDING OF STOCHASTIC OPTIMIZATION IN DEEP LEARNING

Our work on the long-run behavior of stochastic gradient descent on non-convex landscapes has opened the door to a principled understanding of the training dynamics of deep learning models. It is now possible to dream of a “grand unified theory” of non-convex optimization for statistical learning: a theoretical framework that would characterize how the geometry of the model, the optimization algorithm and the distribution of the data interact to shape the training dynamics and the generalization performance of the end model.

More precisely, the fundamental components that need to be developed to build this theory are:

- *Extending our framework to adaptive methods such as AdaGrad (Duchi et al., 2011) and Adam (Kingma and Ba, 2014):* the behavior of these methods is understood even less than that of SGD, especially in the non-convex setting, yet their empirical performance is often superior to that of SGD. Understanding why this is the case would be a major step forward. From a technical point of view, this would require adapting the proof techniques of Chapter 3, since our current framework heavily relies on the fact that SGD is a noisy discretization of a gradient flow. An ODE approximation for Adam in the spirit of (Ma et al., 2022, Barakat et al., 2021, Dereich et al., 2025) would be a natural starting point. Some preliminary results in the analysis of the global convergence time for Adam are plotted in Figure 25.
- *Refining our theory to understand the implicit bias of gradient methods:* for now, the theory of Chapter 4 describes which components of the set of critical points are selected by the long-run dynamics of SGD, but it remains agnostic to the behaviour of SGD inside these components.

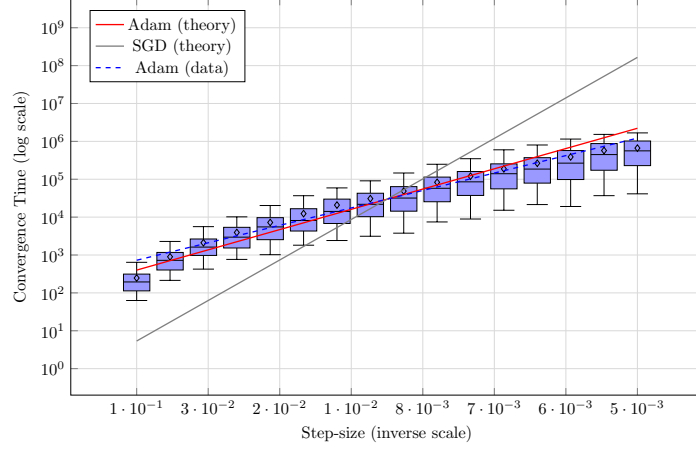


Figure 25: Global convergence time of multiple runs of Adam on a 2D non-convex problem with Gaussian noise in the setting of § 5.1. The empirical results for Adam are in blue, theoretical prediction for Adam in red and, for comparison, theoretical prediction for SGD in gray.

This is particularly relevant in the overparameterized regime: when the model is sufficiently overparameterized, the set of critical points contains large connected components, and the implicit bias of SGD within these components remains a key open question (Li et al., 2021).¹

- *Training dynamics and generalization:* a theoretical understanding of the generalization of deep learning models has remained elusive, in part because of the lack of a precise description of the training dynamics. We can now tackle this question by building on our framework. A natural starting point for this would be to understand how the invariant measures of SGD running on the training data relate to the invariant measures of SGD running on the population distribution, taking inspiration from Raginsky and Recht (2026). This stability analysis would shed further light on how the training dynamics shape the generalization as well as robustness and privacy in the trained model.
- *Going beyond minimization problems:* many training procedures in deep learning aim not only to minimize a loss function but to solve even more intricate problems. Reinforcement learning (RL) is a foremost example: training dynamics in RL are notoriously unstable and difficult to tune. Extending our framework to these settings would provide a principled understanding of these dynamics and inform the design of more reliable training algorithms.²
- *Understanding how the design of deep learning architectures influences the efficiency of optimization algorithms:* architectural innovations are typically discovered empirically and then studied primarily through expressivity, while their optimization implications are much less understood. A key example is the attention mechanism: beyond its approximation benefits (Yun et al., 2020), it seems to make training dramatically easier in practice, yet we still lack a grounded explanation for these gains. By identifying the

¹ This is a collaborative project with Tùng Lê Quốc.

² Some preliminary results have been obtained in collaboration with fellow PhD student Pierre-Louis Cauvin.

key geometric quantities that drive non-convex convergence of gradient methods—through exact characterizations of convergence time rather than loose upper bounds—our framework provides a principled basis for studying this phenomenon. Obtaining a clear picture in realistic regimes (finite width and depth) would provide foundational insight into how architectural choices shape training dynamics, and more broadly into the interplay between model design and optimization.

8.2 BUILDING MACHINE LEARNING SYSTEMS THAT WE CAN TRUST

My second long-term goal is to address the robustness and reliability issues that plague modern machine learning systems. Deep learning systems are indeed often brittle, making their deployment challenging. A case in point is LLM systems, which appear overconfident in their predictions even when they are wrong. I propose to address this problem with several key approaches in the short term, building on my previous work on robust optimization and min-max optimization ([Chapter 6](#) and [Chapter 7](#)):

- *Adaptive conformal prediction*: conformal prediction has emerged as a promising paradigm to provide reliable uncertainty quantification, even for large underlying models. This approach is particularly suited for LLMs: it is training-agnostic and computationally cheap at inference time — see e.g., the recent review ([Campos et al., 2024](#)). Conformal prediction can also robustify uncertainty quantification methods that are sensitive to topic drift ([Azizian et al., 2025b](#)). However, existing methods are overly pessimistic, overestimating the uncertainty. I plan to develop refined conformal prediction methods that produce adaptive estimates in the spirit of ([Kiyani et al., 2024](#), [Cherian et al., 2024](#), [Braun et al., 2025](#)) while retaining the same theoretical guarantees that make conformal prediction appealing.
- *WDRO with heavy-tailed distributions*: most statistical perspectives on distributionally robust optimization (DRO) assume that the underlying data distribution has light tails, while in several applications — natural language processing, finance, energy,... — “bad” events are often driven by heavy-tailed distributions. Investigating the properties of DRO models in this setting is thus an interesting avenue for future research.
- *Risk-sensitive reinforcement learning (RL)*: to be deployed in critical applications, RL systems must control downside risk rather than optimize expected reward alone. Tail-risk objectives such as CVaR [Rockafellar and Uryasev \(2000\)](#) are difficult to optimize from data because they are driven by rare trajectories, making standard estimators sample-inefficient and often unstable at scale ([Chow et al., 2018](#)). I intend to develop scalable risk-sensitive RL methods based on a fixed-point perspective on risk measures ([Newey and Powell, 1987](#), [Ziegel, 2016](#)), leading to online estimators that reuse information across iterations while remaining grounded in fresh minibatches of rollouts.
- *Improved min-max optimization algorithms for robust optimization*: most, if not all, robust optimization formulations in deep learning lead to min-max problems. However, we still lack reliable optimization procedures to solve these problems at scale. I plan to tackle this challenge by exploring several avenues: (i) introducing an entropic smoothing of the inner maximization problem, inspired by our work on WDRO ([Azizian et al., 2023b](#)) and developing particle-based methods to solve the resulting smoothed problems

efficiently at scale; *(ii)* studying and developing bespoke regularization terms, in the spirit of the gradient penalty approach, ([Mescheder et al., 2017](#)), whose benefits appear both in practice ([Huang et al., 2024](#)) and in theory ([Azizian et al., 2020b,a](#)).

The long-term ambition is a principled framework for trustworthy machine learning, weaving together robust training and inference procedures.

BIBLIOGRAPHY

- Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. *The Journal of Machine Learning Research*, 22(1):5662–5720, 2021. (cited on pp. 288, 300, and 312)
- Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding adam optimizer via online learning of updates: Adam is FTRL in disguise. In *Forty-first International Conference on Machine Learning, ICML 2024, Proceedings of Machine Learning Research*, pages 619–640. PMLR, 2024. URL <https://proceedings.mlr.press/v235/ahn24b.html>. (cited on p. 68)
- Ahmet Alacaoglu, Yura Malitsky, and Stephen J Wright. Towards weaker variance assumptions for stochastic optimization. *arXiv preprint arXiv:2504.09951*, 2025. (cited on p. 27)
- Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization, June 2019. (cited on p. 19)
- John M. Alongi and Gail Susan Nelson. *Recurrence and Topology*. American Mathematical Soc., 2007. (cited on p. 187)
- Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, 43(2):477–501, 2004. (cited on pp. 119 and 131)
- Yang An and Rui Gao. Generalization bounds for (Wasserstein) robust optimization. *Advances in Neural Information Processing Systems*, 34, 2021. (cited on pp. 97, 98, 99, 101, and 282)
- Kimon Antonakopoulos, E. Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox algorithm for variational inequalities with singular operators. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019. (cited on p. 144)
- Michael Arbel and Julien Mairal. Non-Convex Bilevel Games with Critical Point Selection Maps. *Advances in Neural Information Processing Systems*, 36:1–34, 2022. (cited on pp. 99, 107, 313, and 321)
- Yossi Arjevani and Michael Field. Annihilation of spurious minima in two-layer relu networks. *Advances in Neural Information Processing Systems*, 35:37510–37523, 2022. (cited on p. 19)
- Adriano Arrigo, Christos Ordoudis, Jalal Kazempour, Zacharie De Grève, Jean-François Toubeau, and François Vallée. Wasserstein distributionally robust chance-constrained optimization for energy and reserve dispatch: An exact and physically-bounded formulation. *European Journal of Operational Research*, 296(1):304–322, 2022. (cited on p. 96)
- Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rkgKBhA5Y7>. Poster. (cited on p. 68)
- Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. (cited on p. 114)
- Alfred Auslender and Marc Teboulle. Interior projection-like methods for monotone variational inequalities. *Mathematical Programming*, 104:39–68, 2005. (cited on p. 114)
- Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic Mirror Descent on Overparameterized Nonlinear Models. *IEEE Trans. Neural Netw. Learning Syst.*, 33(12):7717–7727,

- December 2022. ISSN 2162-237X, 2162-2388. doi: [10.1109/TNNLS.2021.3087480](https://doi.org/10.1109/TNNLS.2021.3087480). (cited on p. 19)
- Waïss Azizian and Marc Lelarge. Expressive power of invariant and equivariant graph neural networks. In *ICLR*, 2021. (cited on p. 8)
- Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *AISTATS*, 2020a. (cited on pp. 8 and 148)
- Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *AISTATS*, 2020b. (cited on pp. 8 and 148)
- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *COLT 2021*, 2021a. (cited on pp. 6 and 7)
- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *COLT*, 2021b. (cited on pp. 113 and 363)
- Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for Wasserstein distributionally robust optimization. *ESAIM: COCV*, 2023a. (cited on pp. 6, 7, 95, 97, 103, 104, 105, 281, and 315)
- Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for Wasserstein Distributionally Robust Optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 2023b. (cited on p. 147)
- Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Exact Generalization Guarantees for (Regularized) Wasserstein Distributionally Robust Models. In *NeurIPS*, 2023c. (cited on pp. 6, 7, 95, and 313)
- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The rate of convergence of Bregman proximal methods: Local geometry vs. regularity vs. sharpness. *SIAM Journal on Optimization*, September 2024a. (cited on pp. 6 and 7)
- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. What is the long-run distribution of stochastic gradient descent? A large deviations analysis. In *ICML 2024*, 2024b. (cited on pp. 5, 6, 229, 233, 234, 235, 236, 245, 261, 262, 268, 272, 279, and 280)
- Waïss Azizian, Franck Iutzeler, Jerome Malick, and Panayotis Mertikopoulos. What is the Long-Run Distribution of Stochastic Gradient Descent? A Large Deviations Analysis. In *ICML*, 2024c. (cited on pp. 5, 21, 51, and 169)
- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The rate of convergence of bregman proximal methods: Local geometry versus regularity versus sharpness. *SIAM Journal on Optimization*, 2024d. (cited on pp. 113 and 357)
- Waïss Azizian, Franck Iutzeler, Jerome Malick, and Panayotis Mertikopoulos. The global convergence time of stochastic gradient descent in non-convex landscapes: Sharp estimates via large deviations. In *ICML*, 2025a. (cited on pp. 5, 6, 21, 71, and 229)
- Waïss Azizian, Michael Kirchhof, Eugene Ndiaye, Louis Béthune, Michal Klein, Pierre Ablin, and marco cuturi. The geometries of truth are orthogonal across tasks. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025b. (cited on pp. 7 and 147)
- Waïss Azizian and Ali Hasan. How does the pretraining distribution shape in-context learning? Task selection, generalization, and robustness. *arXiv preprint arXiv: 2510.01163*, 2025. (cited on p. 7)
- Dragana Bajovic, Dusan Jakovetic, and Soumya Kar. Large deviations rates for stochastic gradient descent with strongly convex functions, November 2022. (cited on p. 20)
- Anas Barakat, Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic optimization with momentum: convergence, fluctuations, and traps avoidance. *Electronic Journal of Statistics*, 15(2):3892–3947, 2021. (cited on p. 145)

- Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. On global convergence of resnets: From finite to infinite width using linear parameterization. *Advances in Neural Information Processing Systems*, 35:16385–16397, 2022. (cited on p. 20)
- Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. Understanding the training of infinitely deep and wide resnets with conditional optimal transport. *Communications on Pure and Applied Mathematics*, 78(11):2149–2205, 2025. (cited on p. 20)
- Manon Baudel, Arnaud Guyader, and Tony Lelièvre. On the hill relation and the mean reaction time for metastable processes. *Stochastic Processes and their Applications*, 155:393–436, 2023. (cited on p. 18)
- Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, May 2017. (cited on pp. 118 and 144)
- Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. (cited on pp. 114, 117, 118, 130, and 358)
- Aharon Ben-Tal and Arkadi Semen Nemirovski. *Lectures on modern convex optimization: Analysis, algorithms, and engineering applications*. MOS-SIAM Series on Optimization. SIAM, 2001. (cited on p. 361)
- Michel Benaïm. Dynamics of stochastic approximation algorithms. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, editors, *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer Berlin Heidelberg, 1999. (cited on p. 16)
- Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000. (cited on p. 16)
- Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020. (cited on p. 16)
- Jose Blanchet and Yang Kang. Semi-supervised learning based on distributionally robust optimization. *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods*, 5:1–33, 2020. (cited on pp. 105, 283, and 290)
- Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. (cited on pp. 101, 285, and 315)
- Jose Blanchet and Alexander Shapiro. Statistical limit theorems in distributionally robust optimization. *arXiv preprint arXiv:2303.14867*, 2023. (cited on pp. 97 and 99)
- Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019. (cited on p. 283)
- Jose Blanchet, Karthyek Murthy, and Viet Anh Nguyen. Statistical analysis of wasserstein distributionally robust estimators. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 227–254. INFORMS, 2021a. (cited on p. 96)
- Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 2021b. (cited on p. 288)
- Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in wasserstein distributionally robust estimation. *Biometrika*, 109(2):295–315, 2022a. (cited on pp. 97, 98, 99, and 100)
- Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 47(2):1500–1529, 2022b. (cited on p. 110)
- Vladimir I. Bogachev. *Measure Theory Volume 1*. Springer, 1 edition, 2007. (cited on p. 299)

- Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press and Hindustan Book Agency, 2008. (cited on p. 16)
- Radu Ioan Bot, Sorin-Mihai Grad, and Gert Wanka. *Duality in Vector Optimization*. Vector Optimization. Springer Berlin Heidelberg, 2009. (cited on p. 284)
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. (cited on pp. 96, 353, and 354)
- Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514 – 560, 2005. (cited on p. 290)
- Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. (cited on p. 322)
- Sacha Braun, Liviu Aolaritei, Michael I Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *arXiv preprint arXiv:2503.19068*, 2025. (cited on p. 147)
- Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. (cited on p. 113)
- Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York, 2010. (cited on p. 286)
- Lawrence D. Brown. Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *Lecture Notes-Monograph Series*, 9:i–279, 1986. (cited on p. 189)
- Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–358, 2015. (cited on pp. 114 and 139)
- Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024. doi: [10.1162/tac1_a_00715](https://doi.org/10.1162/tac1_a_00715). URL <https://aclanthology.org/2024.tac1-1.82/>. (cited on p. 147)
- Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017. (cited on pp. 105, 281, 282, and 293)
- O. Catoni. Sharp large deviations estimates for simulated annealing algorithms. *Annales De L Institut Henri Poincare-probabilites Et Statistiques*, 1991. (cited on p. 20)
- Olivier Catoni. Rough Large Deviation Estimates for Simulated Annealing: Application to Exponential Schedules. *The Annals of Probability*, 20(3):1109–1146, July 1992. (cited on p. 20)
- Raphaël Cerf. The dynamics of mutation-selection algorithms with large population sizes. *Annales de l’I.H.P. Probabilités et statistiques*, 32(4):455–508, 1996. (cited on p. 20)
- Raphaël Cerf. Cours “Systèmes dynamiques”. Course lecture, 2019. Oral lecture, Département de Mathématiques et Applications, École Normale Supérieure de Paris, Fall 2019. (cited on p. 52)
- Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, August 1993. (cited on p. 113)
- Ruidi Chen and Ioannis C Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13), 2018. (cited on pp. 97 and 107)
- Ruidi Chen and Ioannis Ch Paschalidis. Distributionally Robust Learning. *Found. Trends Optim.*, 4:1–243, 2020. (cited on p. 282)

- Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting, July 2020. (cited on p. 18)
- John Cherian, Isaac Gibbs, and Emmanuel Candes. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024. (cited on p. 147)
- Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT '12: Proceedings of the 25th Annual Conference on Learning Theory*, 2012. (cited on p. 117)
- Lenaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018. (cited on p. 19)
- Lenaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pages 1305–1338. PMLR, 2020. (cited on p. 20)
- Lenaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019. (cited on p. 20)
- Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018. URL <http://jmlr.org/papers/v18/15-636.html>. (cited on p. 147)
- Constantin Christof and Julia Kowalczyk. On the Omnipresence of Spurious Local Minima in Certain Neural Network Training Problems. *Constr Approx*, June 2023. ISSN 1432-0940. doi: [10.1007/s00365-023-09658-w](https://doi.org/10.1007/s00365-023-09658-w). (cited on p. 19)
- Kuo-Liang Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954. (cited on pp. 364 and 365)
- Christian Clason and Tuomo Valkonen. Introduction to nonsmooth analysis and optimization. *arXiv:2001.00216*, 2020. (cited on p. 284)
- Michel Coste. An introduction to o-minimal geometry. November 1999. (cited on pp. 26, 170, and 232)
- Shisheng Cui and Uday V. Shanbhag. On the analysis of reflected gradient and splitting methods for monotone stochastic variational inequality problems. In *CDC '16: Proceedings of the 57th IEEE Annual Conference on Decision and Control*, 2016. (cited on p. 141)
- Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. (cited on p. 281)
- Léo Dana, Francis Bach, and Loucas Pillaud-Vivien. Convergence of shallow relu networks on weakly interacting data. *arXiv preprint arXiv:2502.16977*, 2025. (cited on p. 20)
- Charles Dapogny, Franck Iutzeler, Andrea Meda, and Boris Thibert. Entropy-regularized wasserstein distributionally robust shape and topology optimization. *Structural and Multidisciplinary Optimization*, 66(3):42, 2023. (cited on p. 105)
- Charles Dapogny, Julien Prando, and Boris Thibert. Distributionally robust shape and topology optimization. *arXiv preprint arXiv: 2507.21574*, 2025. (cited on p. 105)
- Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS '19: Proceedings of the 10th Conference on Innovations in Theoretical Computer Science*, 2019. (cited on p. 141)
- Alejandro D. de Acosta. *Large Deviations for Markov Chains*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 2022. (cited on p. 20)
- Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*, 2023. (cited on p. 68)

- Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. *Neural Information Processing Systems*, 2024. doi: [10.48550/arXiv.2405.15682](https://doi.org/10.48550/arXiv.2405.15682). (cited on p. 68)
- Erick Delage and Yinyu Ye. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, 58: 595–612, 2010. (cited on p. 96)
- Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer, Berlin, Heidelberg, 2010. (cited on p. 20)
- Steffen Dereich, Arnulf Jentzen, and Sebastian Kassing. Ode approximation for the adam algorithm: General and overparametrized setting. *arXiv preprint arXiv:2511.04622*, 2025. (cited on p. 145)
- Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3), 2020. (cited on p. 18)
- Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham, 2018. (cited on pp. 35, 55, 193, 201, 206, 216, 217, and 238)
- Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019. (cited on p. 19)
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12: 2121–2159, 2011. (cited on p. 145)
- Paul Dupuis. Large Deviations Analysis of Some Recursive Algorithms with State Dependent Noise. *The Annals of Probability*, 16(4):1509–1536, October 1988. (cited on p. 20)
- Paul Dupuis and Harold J. Kushner. Stochastic Approximations via Large Deviations: Asymptotic Properties. *SIAM J. Control Optim.*, 23(5):675–696, September 1985. (cited on p. 20)
- Paul Dupuis and Harold J. Kushner. Stochastic Approximation and Large Deviations: General Results for W.p.l. Convergence,;. Technical report, Defense Technical Information Center, Fort Belvoir, VA, February 1987. (cited on p. 20)
- Andreas Eberle. Reflection coupling and Wasserstein contractivity without convexity. *Comptes Rendus Mathématique*, 349(19):1101–1104, November 2011. (cited on p. 18)
- Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probab. Theory Relat. Fields*, 166(3-4):851–886, December 2016. (cited on p. 18)
- Andreas Eberle and Mateusz B. Majka. Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24(none):1–36, January 2019. (cited on p. 18)
- Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Quantitative Harris type theorems for diffusions and McKean-Vlasov processes, October 2017. (cited on p. 18)
- Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, July 2019. (cited on p. 18)
- Murat A. Erdogdu, Lester W. Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. (cited on pp. 18, 28, and 66)
- Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018. (cited on pp. 96, 97, 101, 109, 281, and 282)

- Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003. (cited on p. 114)
- Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trounev, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019. (cited on p. 346)
- Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015. (cited on p. 96)
- M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260 of *Grundlehren Der Mathematischen Wissenschaften*. Springer, New York, NY, 1998a. (cited on p. 20)
- Mark Iosifovich Freidlin and Alexander Dmitrievich Wentzell. *Random Perturbations of Dynamical Systems*. Springer, 1 edition, 1998b. (cited on pp. 12, 18, 21, 22, 23, 24, 29, 31, 32, 33, 37, 38, 46, 47, 48, 51, 53, 54, 76, 174, 175, 176, 177, 200, 201, 205, 206, 211, 219, 220, 229, 230, 234, 236, 237, 238, 242, 245, 246, 250, and 272)
- Drew Fudenberg and Jean Tirole. *Game Theory*. The MIT Press, 1991. (cited on p. 116)
- Arnab Ganguly. Large deviation principle for stochastic integrals and stochastic differential equations driven by infinite-dimensional semimartingales. *Stochastic Processes and their Applications*, 128(7):2179–2227, 2018. (cited on p. 20)
- Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 2022. (cited on pp. 97, 99, and 101)
- Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 2022. (cited on pp. 101, 285, 288, and 315)
- Timur Garipov, Pavel Izmailov, Dmitrii Podoprikin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada*, pages 8803–8812, 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/be3087e74e9100d4bc4c6268cdbe8456-Abstract.html>. (cited on p. 68)
- Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023. (cited on p. 27)
- Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – Online stochastic gradient for tensor decomposition. In *COLT ’15: Proceedings of the 28th Annual Conference on Learning Theory*, 2015a. (cited on p. 57)
- Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. *COLT*, 2015b. (cited on p. 85)
- Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Thirtieth Annual Conference on Neural Information Processing System*, Advances in Neural Information Processing Systems, 2016. (cited on pp. 281, 289, and 346)
- Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019. (cited on pp. 105, 106, 281, 282, and 293)
- Giacomo Di Gesù, Tony Lelièvre, Dorian Le Peutrec, and Boris Nectoux. Sharp asymptotics of the first exit point density. *Annals of PDE*, 5:1–174, 2019. (cited on p. 18)

- Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019. (cited on pp. 115, 140, and 141)
- Joel Goh and Melvyn Sim. Distributionally Robust Optimization and Its Tractable Approximations. *Operations Research*, 58:902–917, 2010. (cited on p. 96)
- Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020a. (cited on p. 140)
- Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *COLT '20: Proceedings of the 33rd Annual Conference on Learning Theory*, 2020b. (cited on p. 140)
- Robert M Gower, Peter Richtárik, and Francis Bach. Stochastic quasi-gradient methods: Variance reduction via jacobian sketching. *Mathematical Programming*, 188(1): 135–192, 2021. (cited on p. 26)
- Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/qian19b.html>. (cited on p. 26)
- Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. (cited on p. 62)
- O. V. Gulinsky and A. Yu. Veretennikov. *Large Deviations for Discrete-Time Processes with Averaging*. De Gruyter, December 1993. (cited on p. 20)
- Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning (ICML)*, pages 3964–3975. PMLR, 2021. (cited on p. 18)
- Elad Hazan. A survey: The convex optimization approach to regret minimization. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*, pages 287–304. MIT Press, 2012. (cited on p. 139)
- Onésimo Hernández-Lerma and Jean Bernard Lasserre. *Markov Chains and Invariant Probabilities*. Birkhäuser, Basel, 2003. (cited on pp. 192 and 193)
- Terrell L Hill. Studies in irreversible thermodynamics iv. diagrammatic representation of steady state fluxes for unimolecular systems. *Journal of theoretical biology*, 10(3): 442–459, 1966. (cited on pp. 51 and 52)
- Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001. (cited on p. 130)
- Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning (ICML)*, pages 4262–4274. PMLR, 2021. (cited on p. 18)
- Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021. (cited on p. 57)
- Ya-Ping Hsieh, Mohammad Reza Karimi, Andreas Krause, and Panayotis Mertikopoulos. Riemannian stochastic optimization methods avoid strict saddle points. In *NeurIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023. (cited on p. 57)

- Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6936–6946, 2019. (cited on pp. 115, 140, and 141)
- Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020. (cited on pp. 142, 371, 374, and 375)
- Daniel Hsu, Sham M Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012. (cited on p. 28)
- Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, 4(1):3–32, 2019. (cited on pp. 18 and 64)
- Yiwen Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The gan is dead; long live the gan! a modern gan baseline. *Advances in Neural Information Processing Systems*, 37:44177–44215, 2024. (cited on p. 148)
- Henrik Hult, Adam Lindhe, Pierre Nyquist, and Guo-Jhen Wu. A weak convergence approach to large deviations for stochastic approximations. 2023. (cited on p. 20)
- Adam Ibrahim, Waiss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International conference on machine learning*, pages 4583–4593. PMLR, 2020. (cited on p. 8)
- Aleksandr Davidovich Ioffe, V. M. Tikhomirov, and K. Makowski. *Theory of Extremal Problems*. North-Holland Pub. Co. ; Sole distributors for the U.S.A. and Canada, Elsevier North-Holland, Amsterdam, New York, 1979. (cited on p. 176)
- Rustem Islamov, Niccolò Ajroldi, Antonio Orvieto, and Aurelien Lucchi. Loss landscape characterization of neural networks without over-parametrization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=h0a3p5WtXU>. (cited on p. 19)
- Alfredo N. Iusem, Alejandro Jofré, Roberto I. Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017. (cited on p. 141)
- Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. In *Conference on Uncertainty in Artificial Intelligence*, 2018. (cited on p. 68)
- Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017. (cited on pp. 18 and 64)
- Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017. URL <http://proceedings.mlr.press/v70/jin17a.html>. (cited on p. 57)
- Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998. (cited on p. 16)
- Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011. (cited on pp. 117, 124, 141, 358, and 363)
- Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13:1865–1890, 2012. (cited on p. 114)
- O. Kallenberg. *Random Measures, Theory and Applications*. Probability Theory and Stochastic Modelling. Springer International Publishing, 2017. (cited on p. 305)

- Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer International Publishing, 2021. (cited on pp. 65, 216, and 217)
- Nigel J. Kalton, Newton T. Peck, and James W. Roberts. *An F-space Sampler*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1984. (cited on p. 299)
- Kenji Kawaguchi. Deep learning without poor local minima. In *NIPS '16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016. (cited on p. 19)
- Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *Transactions on Machine Learning Research*, 2023. (cited on p. 26)
- Rafail Khasminskii. *Stochastic Stability of Differential Equations*, volume 66 of *Stochastic Modelling and Applied Probability*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. (cited on p. 55)
- Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952. (cited on p. 16)
- Yuri Kifer. *Random Perturbations of Dynamical Systems*. Birkhäuser, Boston, MA, 1988a. (cited on pp. 12, 21, 22, 23, 24, 29, 32, 35, 37, 38, 40, 44, 46, 47, 48, 181, 182, 185, 195, 200, and 201)
- Yuri Kifer. *Random Perturbations of Dynamical Systems*. Birkhäuser Boston, Boston, MA, 1988b. (cited on pp. 20, 235, 236, 237, and 238)
- Yuri Kifer. A Discrete-Time Version of the Wentzell-Friedlin Theory. *The Annals of Probability*, 18(4):1676–1692, October 1990. (cited on p. 55)
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014. (cited on p. 145)
- Krzysztof C. Kiwiel. Proximal minimization methods with generalized Bregman functions. *SIAM Journal on Control and Optimization*, 35:1142–1168, 1997. (cited on p. 113)
- Shayan Kiyani, George J Pappas, and Hamed Hassani. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563, 2024. (cited on p. 147)
- Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, 2014. (cited on p. 308)
- G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976. (cited on pp. 114 and 139)
- Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*. INFORMS, 2019. (cited on pp. 96, 281, and 282)
- Yongchan Kwon, Wonyoung Kim, Joong-Ho Won, and Myunghee Cho Paik. Principled learning method for Wasserstein distributionally robust optimization with local perturbations. In *International Conference on Machine Learning*, pages 5567–5576. PMLR, 2020. (cited on p. 96)
- Yassine Laguel, Jérôme Malick, and Zaid Harchaoui. First-order optimization for superquantile-based supervised learning. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020. (cited on p. 96)
- Andrew Lamperski. Projected Stochastic Gradient Langevin Algorithms for Constrained Sampling and Non-Convex Learning. In *Conference on Learning Theory (COLT)*, pages 2891–2937. PMLR, July 2021. (cited on p. 18)
- Lev Davidovich Landau and Evgeny Mikhailovich Lifshitz. Statistical physics. In *Course of Theoretical Physics*, volume 5. Pergamon Press, Oxford, 1976. (cited on p. 56)

- Tam Le and Jerome Malick. Universal generalization guarantees for wasserstein distributionally robust models. In *The Thirteenth International Conference on Learning Representations*, 2025. (cited on p. 103)
- Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018. (cited on pp. 97 and 98)
- John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018. (cited on pp. 321 and 322)
- Se Yoon Lee. Gibbs sampler and coordinate ascent variational inference: A set-theoretical review. *Communications in Statistics - Theory and Methods*, pages 1–21, 2021. (cited on p. 290)
- FT Leighton and RL Rivest. Estimating a probability using finite memory. *IEEE Transactions on Information Theory*, 32(6):733–742, 1986. (cited on pp. 51 and 52)
- Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33, 2020. (cited on p. 96)
- Dawei Li, Tian Ding, and Ruoyu Sun. On the benefit of width for neural networks: Disappearance of basins. *SIAM Journal on Optimization*, 32(3):1728–1758, 2022a. (cited on p. 19)
- Jiajin Li, Caihua Chen, and Anthony Man-Cho So. Fast Epigraphical Projection-based Incremental Algorithms for Wasserstein Distributionally Robust Support Vector Machine. In *Advances in Neural Information Processing Systems*, volume 33, pages 4029–4039. Curran Associates, Inc., 2020a. (cited on p. 96)
- Lei Li and Yuliang Wang. A sharp uniform-in-time error estimate for Stochastic Gradient Langevin Dynamics. *arXiv preprint arXiv:2207.09304*, 2022. (cited on pp. 18, 28, and 66)
- Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning (ICML)*, pages 2101–2110. PMLR, 2017. (cited on p. 16)
- Qianxiao Li, Cheng Tai, and Weinan E. Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019. (cited on p. 16)
- Xiang Li, Zebang Shen, Liang Zhang, and Niao He. A hessian-aware stochastic differential equation for modelling sgd. *ArXiv*, abs/2405.18373, 2024a. (cited on p. 17)
- Xiang Li et al. Minima selection in stochastic optimization: A long-time, small-stepsizes perspective. *Preprint*, 2024b. URL https://odi.inf.ethz.ch/papers/Minima_Selection_long_time_small_stepsizes_perspective.pdf. (cited on p. 17)
- Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14544–14555, 2020b. (cited on pp. 17 and 28)
- Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss?—a mathematical framework. In *International Conference on Learning Representations (ICLR)*, 2021. (cited on pp. 16, 17, and 146)
- Zhiyuan Li, Tianhao Wang, and Dingli Yu. Fast mixing of stochastic gradient descent with normalization and weight decay. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 9233–9248, 2022b. (cited on p. 17)
- Chaoyue Liu, Dmitriy Drusvyatskiy, Mikhail Belkin, Damek Davis, and Yi-An Ma. Aiming towards the minimizers: Fast convergence of SGD for overparametrized problems, June 2023. (cited on p. 19)
- Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4):551–575, August 1977. (cited on p. 16)
- Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively-smooth convex optimization by first-order methods and applications. *SIAM Journal on Optimization*, 28(1): 333–354, 2018. (cited on p. 144)

- Yu Lu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. (cited on pp. 18, 28, and 66)
- Aurelien Lucchi, Frank Proske, Antonio Orvieto, Francis Bach, and Hans Kersting. On the theoretical properties of noise correlation in stochastic optimization. *Advances in Neural Information Processing Systems*, 35:14261–14273, 2022. (cited on p. 18)
- Alessandra Lunardi. *Interpolation theory*, volume 9. Springer, 2009. (cited on p. 302)
- Iosif Lytras and Panayotis Mertikopoulos. Tamed Langevin sampling under weaker conditions. <https://arxiv.org/abs/2405.17693>, 2024. (cited on p. 28)
- Chao Ma, Lei Wu, et al. A qualitative study of the dynamic behavior for adaptive gradient algorithms. In *Mathematical and scientific machine learning*, pages 671–692. PMLR, 2022. (cited on p. 145)
- Mateusz B. Majka, Aleksandar Mijatović, and Łukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, 30(4):1534 – 1581, 2020. (cited on p. 18)
- Yura Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015. (cited on pp. 115 and 140)
- Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. *Advances in neural information processing systems*, 36:63210–63221, 2023. (cited on p. 20)
- Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Transformative or conservative? conservation laws for resnets and transformers. In *International Conference on Machine Learning*, pages 43140–43176. PMLR, 2025. (cited on p. 20)
- Pierre Marion, Yu-Han Wu, Michael Eli Sander, and Gérard Biau. Implicit regularization of deep residual networks towards neural ODEs. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=AbXGwqb5Ht>. (cited on p. 19)
- Simon Martin, Giulio Biroli, and Francis Bach. High-dimensional analysis of gradient flow for extensive-width quadratic neural networks. *arXiv preprint arXiv:2601.10483*, 2026. (cited on p. 17)
- Quentin Merigot and Boris Thibert. Optimal transport: discretization and algorithms. In *Handbook of Numerical Analysis*, volume 22, pages 133–212. Elsevier, 2021. (cited on p. 281)
- Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, November 2016. (cited on p. 119)
- Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, January 2018a. (cited on p. 16)
- Panayotis Mertikopoulos and Mathias Staudigl. Stochastic mirror descent dynamics and their convergence in monotone variational inequalities. *Journal of Optimization Theory and Applications*, 179(3):838–867, December 2018b. (cited on p. 16)
- Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2): 465–507, January 2019. (cited on p. 358)
- Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR ’19: Proceedings of the 2019 International Conference on Learning Representations*, 2019. (cited on pp. 124, 358, 359, and 363)
- Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. In

- NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020. (cited on pp. 16, 28, and 57)
- Panayotis Mertikopoulos, Ya-Ping Hsieh, and Volkan Cevher. A unified stochastic approximation framework for learning in games. *Mathematical Programming*, 203: 559–609, January 2024. (cited on p. 16)
- Lars M. Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017*. (cited on p. 148)
- Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, London, 1993. (cited on p. 55)
- Francesca Mignacco and Pierfrancesco Urbani. The effective noise of stochastic gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083405, 2022. (cited on p. 17)
- Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9540–9550, 2020. (cited on p. 17)
- Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. <https://arxiv.org/abs/1901.08511v2>, 2019a. (cited on p. 140)
- Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. <https://arxiv.org/pdf/1906.01115.pdf>, 2019b. (cited on p. 140)
- Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. In *AISTATS '20: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020. (cited on p. 115)
- Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of sgd. In *International Conference on Machine Learning (ICML)*, pages 15959–15975. PMLR, 2022. (cited on pp. 18, 59, 61, 64, 65, 87, and 221)
- Julian Musielak. *Orlicz spaces and modular spaces*, volume 1034. Springer, 2006. (cited on p. 299)
- Arkadi Semen Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. (cited on pp. 114, 117, and 144)
- Arkadi Semen Nemirovski and David Berkovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY, 1983. (cited on pp. 113, 114, and 117)
- Arkadi Semen Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. (cited on p. 117)
- Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004. (cited on p. 115)
- Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009. (cited on p. 114)
- Whitney K. Newey and James L. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847, 1987. doi: 10.2307/1911031. (cited on p. 147)
- Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International conference on machine learning*, pages 2603–2612. PMLR, 2017. (cited on pp. 19, 85, and 88)

- Quynh Nguyen and Matthias Hein. Optimization landscape and expressivity of deep cnns. In *International conference on machine learning*, pages 3730–3739. PMLR, 2018. (cited on p. 19)
- Quynh Nguyen, Mahesh Chandra Mukkamala, and Matthias Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. In *International Conference on Learning Representations*, 2019a. URL <https://openreview.net/forum?id=HJgXsjA5tQ>. (cited on pp. 19, 85, and 88)
- Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020. (cited on p. 19)
- Quynh N Nguyen, Pierre Br  chet, and Marco Mondelli. When are solutions connected in deep networks? *Advances in Neural Information Processing Systems*, 34:20956–20969, 2021. (cited on pp. 19 and 88)
- Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Ga  l RICHARD. First Exit Time Analysis of Stochastic Gradient Descent Under Heavy-Tailed Gradient Noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019b. (cited on p. 18)
- Evgenii Nikishin, Pavel Izmailov, Ben Athiwaratkun, Dmitrii Podoprikin, Timur Garipov, Pavel Shvechikov, Dmitry Vetrov, and Andrew Gordon Wilson. Improving stability in deep reinforcement learning with weight averaging. In *UAI 2018 Workshop on Uncertainty in Deep Learning*, 2018. URL <https://www.gatsby.ucl.ac.uk/~balaji/udl-camera-ready/UDL-24.pdf>. (cited on p. 68)
- Antonio Orvieto, Hans Kersting, Frank Proske, Francis Bach, and Aurelien Lucchi. Anti-correlated noise injection for improved generalization. In *International Conference on Machine Learning*, pages 17094–17116. PMLR, 2022. (cited on p. 18)
- Antonio Orvieto, Anant Raj, Hans Kersting, and Francis Bach. Explicit regularization in overparametrized models via noise injection. In *International Conference on Artificial Intelligence and Statistics*, pages 7265–7287. PMLR, 2023. (cited on p. 18)
- Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020. (cited on p. 19)
- Fran  ois-Pierre Paty and Marco Cuturi. Regularized Optimal Transport is Ground Cost Adversarial. In *ICML*, 2020a. (cited on p. 346)
- Fran  ois-Pierre Paty and Marco Cuturi. Regularized optimal transport is ground cost adversarial. In *International Conference on Machine Learning*, pages 7532–7542. PMLR, 2020b. (cited on pp. 281, 282, and 285)
- Krunoslav Lehman Pavasovic, Alain Durmus, and Umut Simsekli. Approximate heavy tails in offline (multi-pass) stochastic gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. (cited on p. 18)
- Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *Annals of Probability*, 18(2):698–712, April 1990. (cited on p. 57)
- Teemu Pennanen and Ari-Pekka Perkki  . Topological duals of locally convex function spaces. *Positivity*, 26(1):1–38, 2022. (cited on pp. 300, 301, 302, and 310)
- Romain Petit, Clarice Poon, and Gabriel Peyr  . On the global convergence of gradient descent for wide shallow models with bounded nonlinearities. *arXiv preprint arXiv:2605.10775*, 2026. (cited on p. 20)
- Henning Petzka and Cristian Sminchisescu. Non-attracting Regions of Local Minima in Deep and Wide Neural Networks. *Journal of Machine Learning Research*, 22(143):1–34, 2021. ISSN 1533-7928. (cited on p. 19)
- Juan Peypouquet. *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. SpringerBriefs in Optimization. Springer International Publishing, 2015. (cited on p. 285)

- Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. (cited on pp. 103, 281, and 346)
- William Piat, Jalal Fadili, Frédéric Jurie, and Sébastien da Veiga. Regularized robust optimization with application to robust learning. 2022. (cited on p. 105)
- Samuele Pollaci. Spurious valleys and clustering behavior of neural networks. In *International Conference on Machine Learning*, pages 28079–28099. PMLR, 2023. (cited on p. 19)
- Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046. (cited on p. 67)
- Boris Teodorovich Polyak. *Introduction to Optimization*. Optimization Software, New York, NY, USA, 1987. (cited on pp. 128, 139, 357, 364, and 365)
- Leonid Denisovich Popov. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5): 845–848, 1980. (cited on pp. 114 and 117)
- Maxim Raginsky and Benjamin Recht. Separating geometry from probability in the analysis of generalization. *arXiv preprint arXiv: 2604.19560*, 2026. (cited on p. 146)
- Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory (COLT)*, pages 1674–1703. PMLR, 2017. (cited on pp. 17, 28, and 66)
- Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2013. (cited on pp. 114 and 117)
- Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023. (cited on p. 190)
- Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951. (cited on p. 16)
- R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038. (cited on p. 147)
- R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, 1998. (cited on pp. 285, 286, 291, 295, and 334)
- Ralph Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970. (cited on pp. 132 and 134)
- Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013. doi: 10.1214/ECP.v18-2865. URL <https://doi.org/10.1214/ECP.v18-2865>. (cited on p. 28)
- Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987. (cited on pp. 283, 288, and 319)
- Walter Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1991. (cited on p. 303)
- David Ruppert. Efficient estimators from a slowly convergent robbins-monro process. Technical Report 781, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1988. (cited on p. 67)
- Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, 2015. (cited on pp. 281 and 297)
- Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, and Francis Bach. The Surprising Agreement Between Convex Optimization Theory and Learning-Rate Scheduling for Large Model Training, 2025. (cited on p. 68)

- Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally Robust Logistic Regression. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. (cited on pp. 96 and 101)
- Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via Mass Transportation. *Journal of Machine Learning Research*, 20:1–68, 2019a. (cited on pp. 282 and 283)
- Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via Mass Transportation. *Journal of Machine Learning Research*, 20:1–68, 2019b. (cited on pp. 96, 97, and 107)
- Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011. (cited on p. 118)
- Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. (cited on pp. 96, 97, and 285)
- Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don’t decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018. (cited on p. 62)
- Eilon Solan and Nicolas Vieille. Stochastic games. *Proceedings of the National Academy of Sciences*, 112(45):13743–13746, 2015. (cited on p. 116)
- Jun Song, Niao He, Lijun Ding, and Chaoyue Zhao. Provably convergent policy optimization via metric-aware trust region methods. *arXiv preprint arXiv:2306.14133*, 2023. (cited on p. 105)
- Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on pp. 96 and 97)
- Hoang Tran, Qinzi Zhang, and Ashok Cutkosky. Reevaluating theoretical analysis methods for optimization in deep learning. *arXiv preprint arXiv:2407.01825*, 2024. (cited on p. 69)
- Alain Trounev. Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms. *Annales de l’I.H.P. Probabilités et statistiques*, 32(3):299–348, 1996. (cited on p. 20)
- Constantino Tsallis. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988. (cited on p. 119)
- Koji Tsuda, Gunnar Rätsch, and Manfred K. Warmuth. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *Journal of Machine Learning Research*, 6:995–1018, 2005. (cited on p. 114)
- Lou Van Den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Math. J.*, 84(2), August 1996. (cited on pp. 26, 170, and 232)
- Robert J. Vanderbei, Marc S. Meketon, and Barry A. Freedman. A modification of Karmarkar’s linear programming algorithm. *Algorithmica*, 1(1):395–407, November 1986. (cited on p. 114)
- Rodrigo Veiga, Anastasia Remizova, and Nicolas Macris. Stochastic gradient flow dynamics of test risk and its exact solution for weak features. *arXiv preprint arXiv:2402.07626*, 2024. (cited on p. 17)
- Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133):1–34, 2019. (cited on p. 19)
- Cédric Villani. *Optimal Transport: Old and New*. Springer, 2008. (cited on p. 309)
- Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Soc., 2003. (cited on pp. 281 and 298)
- Florian Vincent, Waïss Azizian, Franck Iutzeler, and Jérôme Malick. skwdro: a library for Wasserstein distributionally robust machine learning. *Journal of Machine Learning Research*, 2026. (cited on pp. 4, 7, and 105)

- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019. (cited on pp. 28, 96, and 321)
- Jie Wang and Yao Xie. A data-driven approach to robust hypothesis testing using sinkhorn uncertainty sets. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 3315–3320. IEEE, 2022. (cited on p. 105)
- Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv:2109.11926*, 2021. (cited on pp. 97, 104, 105, 110, 283, and 290)
- Jie Wang, Ronald Moore, Yao Xie, and Rishikesan Kamaleswaran. Improving sepsis prediction model generalization with optimal transport. In *Machine Learning for Health*, pages 474–488. PMLR, 2022a. (cited on p. 105)
- Xingyu Wang, Sewoong Oh, and Chang-Han Rhee. Eliminating Sharp Minima from SGD with Truncated Heavy-tailed Noise, May 2022b. (cited on p. 18)
- Yi Wang and Zhiren Wang. Three-stage evolution and fast equilibrium for sgd with non-degenerate critical points. In *International Conference on Machine Learning (ICML)*, pages 23092–23113. PMLR, 2022. (cited on p. 17)
- Yifei Wang, Jonathan Lacotte, and Mert Pilanci. The hidden convex optimization landscape of regularized two-layer reLU networks: an exact characterization of optimal solutions. In *International Conference on Learning Representations*, 2022c. URL <https://openreview.net/forum?id=Z7Lk2cQEG8a>. (cited on p. 19)
- Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021. (cited on p. 115)
- E. Weinan, Chao Ma, Qingcan Wang, and Lei Wu. Analysis of the Gradient Descent Algorithm for a Deep Neural Network Model with Skip-connections. *ArXiv*, April 2019. (cited on p. 19)
- Simon Weissmann, Sara Klein, Waïss Azizian, and Leif Döring. Almost sure convergence of stochastic gradient methods under gradient domination. *Transactions on Machine Learning Research*, 2025. (cited on p. 7)
- Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part i: Discrete time analysis. *Journal of Nonlinear Science*, 33(3):45, 2023a. (cited on p. 19)
- Stephan Wojtowytsch. Stochastic Gradient Descent with Noise of Machine Learning Type Part II: Continuous Time Analysis. *J Nonlinear Sci*, 34(1):16, November 2023b. (cited on p. 17)
- Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020. (cited on p. 18)
- Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations (ICLR)*, 2020. (cited on pp. 18, 59, and 64)
- Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Chris De Sa. SWALP: Stochastic weight averaging in low precision training. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7015–7024. PMLR, 2019. URL <https://proceedings.mlr.press/v97/yang19d.html>. (cited on p. 68)
- Jiaojiao Yang, Wenqing Hu, and Chris Junchi Li. On the fast convergence of random perturbations of the gradient flow, April 2020. (cited on p. 16)
- Yaodong Yu, Tianyi Lin, Eric V Mazumdar, and Michael Jordan. Fast distributionally robust learning with variance-reduced min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1219–1250. PMLR, 2022. (cited on p. 282)

- Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020*. (cited on p. 146)
- Constantin Zalinescu. *Convex analysis in general vector spaces*. World scientific, 2002. (cited on pp. 303 and 304)
- Yibo Zeng and Henry Lam. Generalization bounds with minimal dependency on hypothesis class via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 35:27576–27590, 2022. (cited on p. 97)
- Huishuai Zhang, Da Yu, Mingyang Yi, Wei Chen, and Tie-Yan Liu. Stabilize deep resnet with a sharp scaling factor τ . *Machine Learning*, 111(9):3359–3392, 2022a. (cited on p. 19)
- Jingzhao Zhang, Haochuan Li, Suvrit Sra, and Ali Jadbabaie. Neural Network Weights Do Not Converge to Stationary Points: An Invariant Measure Perspective, June 2022b. (cited on p. 67)
- Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with wasserstein metric. *Operations Research Letters*, 46(2):262–267, 2018. (cited on p. 96)
- Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8156–8165, June 2021. (cited on p. 18)
- Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen P. Boyd, and Peter W. Glynn. Stochastic mirror descent for variationally coherent optimization problems. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017. (cited on p. 16)
- Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen P. Boyd, and Peter W. Glynn. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020. (cited on p. 16)
- Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 280–288. PMLR, 2021a. (cited on p. 96)
- Jia-Jie Zhu, Christina Kouridi, Yassine Nemmour, and Bernhard Schölkopf. Adversarially robust kernel smoothing. *arXiv preprint arXiv:2102.08474*, 2021b. (cited on p. 96)
- Johanna F. Ziegel. Coherence and elicibility. *Mathematical Finance*, 26(4):901–918, 2016. doi: [10.1111/mafi.12080](https://doi.org/10.1111/mafi.12080). (cited on p. 147)
- Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in neural information processing systems*, 32, 2019. (cited on p. 19)
- Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020a. (cited on p. 19)
- Difan Zou, Philip M. Long, and Quanquan Gu. On the global convergence of training deep linear resnets. In *International Conference on Learning Representations*, 2020b. URL <https://openreview.net/forum?id=HJxEhREKDH>. (cited on p. 19)
- Bibliographic comments go here, then bibliography.

APPENDICES



WHAT IS THE LONG-RUN DISTRIBUTION OF STOCHASTIC GRADIENT DESCENT? A LARGE DEVIATIONS ANALYSIS

This appendix contains the proofs for [Chapter 4](#) and [Chapter 3](#). It is extracted from [Azizian et al. \(2024c\)](#).

A.1 SETUP AND PRELIMINARIES

A.1.1 Notation

Throughout the sequel, we will write $\langle \cdot, \cdot \rangle$ for the standard inner product on $\mathbb{R}^d \equiv \mathbb{R}^d$ and $\|\cdot\|$ for the induced (Euclidean) norm. To lighten notation, we will identify $\mathbb{R}^d$ with its dual $\mathbb{R}^d \equiv (\mathbb{R}^d)^*$, and we will not formally distinguish between primal and dual vectors (though the distinction should be clear from the context). We will also write $\mathbb{B}(x, r)$ (resp. $\bar{\mathbb{B}}(x, r)$) for the open (resp. closed) ball of radius r centered at $x \in \mathbb{R}^d$, and we will respectively denote the open and closed δ -neighborhoods of $A \subseteq \mathbb{R}^d$ as

$$\mathcal{U}_\delta(A) := \{x \in \mathbb{R}^d : \text{dist}(A, x) < \delta\} \quad (\text{A.1.1a})$$

$$A_\delta := \text{cl}(\mathcal{U}_\delta(A)) = \{x \in \mathbb{R}^d : \text{dist}(A, x) \leq \delta\} \quad (\text{A.1.1b})$$

A.1.2 Setup and assumptions

Before we begin our proof, we revisit and discuss our standing assumptions. In particular, to extend the range of our results, we provide in the rest of this appendix a weaker version of the blanket assumptions of [Appendix A.1.2](#), which we label with an asterisk (“*”) and which will be in force throughout the appendix.

We begin with our assumptions for the objective function f .

Assumption A.1 (Weaker version of [Assumption A.1](#)). The objective function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the following conditions:

- (a) *Coercivity*: $f(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$.
- (b) *Smoothness*: f is C^2 -differentiable and its gradient is β -Lipschitz continuous, that is,

$$\|\nabla f(x') - \nabla f(x)\| \leq \beta \|x' - x\| \quad \text{for all } x, x' \in \mathbb{R}^d. \quad (\text{A.1.2})$$

- (c) *Critical set regularity*: The critical set

$$\text{crit } f := \{x \in \mathbb{R}^d : \nabla f(x) = 0\} \quad (\text{A.1.3})$$

of f consists of a finite number of essentially smoothly connected components $\mathcal{K}_i, i = 1, \dots, K$.

The difference between [Assumption A.1](#) and [Assumption A.1](#) is that, in the latter, the connected components of $\text{crit } f$ are only required to be *essentially* smoothly connected. Formally, this means that, for any connected component $\mathcal{K}$ of $\text{crit } f$, and for any two points $x, x' \in \mathcal{K}$, there exists (a) a continuous, almost everywhere differentiable curve $\gamma: [0, 1] \rightarrow \mathcal{K}$ such that $\gamma(0) = x, \gamma(1) = x'$; and (b) a partition $0 = t_0 < \dots < t_n < \dots < t_N = 1$ of $[0, 1]$ such that γ is integrable on every closed interval of (t_{n-1}, t_n) for all $n = 1, \dots, N$.

Remark 3. The path-connectedness requirement of [Assumption A.1\(c\)](#) is satisfied whenever the connected components of $\text{crit } f$ are isolated critical points, smooth manifolds, or finite unions of closed manifolds. More generally, [Assumption A.1\(c\)](#) is satisfied whenever f is definable – in which case $\text{crit } f$ is also definable, so each component can be connected by piecewise smooth paths ([Coste, 1999](#), [Van Den Dries and Miller, 1996](#)). The relaxation provided by [Assumption A.1\(c\)](#) represents the “minimal” set of hypotheses that are required for our analysis to go through. 

Moving forward, to align our notation with standard conventions in large deviations theory, it will be more convenient to work with $-Z(x; \omega)$ instead of $Z(x; \omega)$ in our proofs. To make this clear, we restate below [Assumption A.2](#) in terms of the noise process

$$U(x, \omega) = -Z(x; \omega) \quad (\text{A.1.4})$$

and we make use of this opportunity to relax the definition of the variance proxy of U .

Assumption A.2 (Weaker version of [Assumption A.2](#)). The noise term $U: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}^d$ satisfies the following properties:

- (a) *Properness*: $\mathbb{E}[U(x, \omega)] = 0$ and $\text{cov}(U(x, \omega)) > 0$ for all $x \in \mathbb{R}^d$.
- (b) *Smooth growth*: $U(x, \omega)$ is C^2 -differentiable and satisfies the growth condition

$$\sup_{x, \omega} \frac{\|U(x, \omega)\|}{1 + \|x\|} < \infty. \quad (\text{A.1.5})$$

- (c) *Sub-Gaussian tails*: The cumulant-generating function of U is bounded as

$$\log \mathbb{E} \left[e^{\langle p, U(x, \omega) \rangle} \right] \leq \frac{1}{2} \sigma_\infty^2(f(x)) \|p\|^2 \quad \text{for all } p \in \mathbb{R}^d, \quad (\text{A.1.6})$$

where $\sigma: \mathbb{R} \rightarrow (0, \infty)$ is continuous, bounded away from zero ($\inf \sigma_\infty^2 > 0$), and grows at infinity as $\sigma_\infty^2(f(x)) = \Theta(\|x\|^s)$ for some $s \in [0, 2]$, i.e.,

$$0 < \liminf_{\|x\| \rightarrow \infty} \frac{\sigma_\infty^2(f(x))}{\|x\|^s} \leq \limsup_{\|x\| \rightarrow \infty} \frac{\sigma_\infty^2(f(x))}{\|x\|^s} < \infty. \quad (\text{A.1.7})$$

Remark 4. The difference between [Assumption A.2](#) and [Assumption A.2](#) lies in the requirements for the variance proxy σ_∞^2 of the noise in [\(A.1.15\)](#). Because f is coercive, the dependence of σ_∞^2 on $f(x)$ allows for noise processes with an unbounded variance as $\|x\| \rightarrow \infty$; the specific functional dependence through $f(x)$ instead of x is a modeling choice which we make because it greatly simplifies the proof and our calculations. 

In this more general setting, we augment [Assumption A.3](#) with a technical condition that is satisfied trivially when σ_∞^2 is constant.

Assumption A.3 (Weaker version of [Assumption A.3](#)). The signal-to-noise ratio of $\mathbf{G}$ is bounded as

$$\liminf_{\|x\| \rightarrow \infty} \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} > 16 \log 6 \cdot d. \quad (\text{A.1.8})$$

Example A.1 (Example of [Appendix A.1.2](#), redux). In [Appendix A.1.2](#), we discussed the regularized empirical risk minimization problem and mentioned that, under a dissipativity condition and with a large enough batch size, it fits our framework. We now provide more details.

Consider the objective f given by

$$f(x) = \frac{1}{N} \sum_{i=1}^N \ell(x; \xi_i) + \frac{\lambda}{2} \|x\|^2 \quad (\text{A.1.9})$$

where $\ell(x; \xi)$ represents the loss of the model x on the data point ξ , ξ_i , $i = 1, \dots, N$, are the training data and λ is the (positive) regularization parameter.

Let us assume that ℓ is non-negative, C^2 -differentiable, β -Lipschitz smooth and that the resulting objective f is dissipative: there are $\alpha, \beta > 0$ such that

$$\langle \nabla f(x), x \rangle \geq \alpha \|x\|^2 - \beta. \quad (\text{A.1.10})$$

As is usually the case in practice, we consider the SFO obtained by sampling mini-batches of size B . The noise term $U(x, \omega)$ is then given by

$$U(x, \omega) = \frac{1}{B} \sum_{b=1}^B \nabla \ell(x; \xi_{\omega_b}) - \frac{1}{N} \sum_{i=1}^N \nabla \ell(x; \xi_i) \quad (\text{A.1.11})$$

with $\omega = (\omega_1, \dots, \omega_B)$ representing B indices from $\{1, \dots, N\}$.

All the terms

$$\nabla \ell(x; \xi_{\omega_b}) - \frac{1}{N} \sum_{i=1}^N \nabla \ell(x; \xi_i) \quad (\text{A.1.12})$$

are uniformly bounded by $\mathcal{O}(\|x\|)$ by smoothness of ℓ . In particular, this implies that [Assumption A.2\(b\)](#) is satisfied.

Moreover, we also obtain that all the terms of the form [\(A.1.12\)](#) are $\mathcal{O}(\|x\|^2)$ -sub-Gaussian, and therefore, by independence we obtain that $U(x, \omega)$ is $\mathcal{O}(\frac{1}{B}\|x\|^2)$ -sub-Gaussian.

Since ℓ is non-negative, f is lower-bounded by $\Omega(\|x\|^2)$. It gives that $U(x, \omega)$ is actually $\mathcal{O}(\frac{1}{B}f(x))$ -sub-Gaussian and therefore [Assumption A.2\(b\)](#) is satisfied with $\sigma_\infty^2(t) \propto \frac{t}{B}$.

We now show that [Assumption A.3](#) can be satisfied by choosing B large enough. Indeed, by dissipativity we have that

$$\begin{aligned} \|\nabla f(x)\|^2 &= \|\nabla f(x) - \alpha x\|^2 + \alpha^2 \|x\|^2 + 2\alpha \langle \nabla f(x) - \alpha x, x \rangle \\ &\geq \alpha^2 \|x\|^2 - 2\beta = \Omega(\|x\|^2), \end{aligned} \quad (\text{A.1.13})$$

so that

$$\frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \geq \Omega\left(\frac{\|x\|^2}{f(x)/B}\right) = \Omega(B). \quad (\text{A.1.14})$$

The second part of [Assumption A.3](#) is satisfied by non-negativity and smoothness of ℓ . 

In this framework, the iterates of (A.1.15), started at $x \in \mathbb{R}^d$, are defined by the following recursion:

$$\begin{cases} x_0 \in \mathbb{R}^d \\ x_{n+1} = x_n - \eta \nabla f(x_n) + \eta U_n, \quad \text{where } U_n = U(x_n, \omega_n) \end{cases} \quad (\text{A.1.15})$$

where $(\omega_n)_{n \geq 0}$ is a sequence of random variables in $\mathbb{R}^m$. We will denote by $\mathbb{P}_x$ the law of the sequence $(\omega_n)_{n \geq 0}$ when the initial point is x and by $\mathbb{E}_x$ the expectation with respect to $\mathbb{P}_x$.

[Assumption A.1](#) and [Assumption A.2](#) imply the following growth condition, that we assume holds with the same constant for the sake of simplicity. There is $M > 0$ such that, for all $x \in \mathbb{R}^d$, $\omega \in \Omega$,

$$\|\nabla f(x)\| \leq M(1 + \|x\|) \quad \text{and} \quad \|U(x, \omega)\| \leq M(1 + \|x\|). \quad (\text{A.1.16})$$

We introduce the cumulant generating functions of the noise $U(x, \omega)$ and of the drift $-\nabla f(x) + U(x, \omega)$, that we denote by $\tilde{\mathcal{H}}, \mathcal{H}$ to avoid confusion. We also define their convex conjugates, that we denote by $\tilde{\mathcal{L}}, \mathcal{L}$.

Definition A.4 (Hamiltonians and Lagrangians). Define, for $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$,

$$\begin{aligned} \tilde{\mathcal{H}}(x, v) &= \log \mathbb{E}[\exp(\langle v, U(x, \omega) \rangle)] \\ \mathcal{H}(x, v) &= -\langle \nabla f(x), v \rangle + \tilde{\mathcal{H}}(x, v) \\ \tilde{\mathcal{L}}(x, v) &= \tilde{\mathcal{H}}(x, \cdot)^*(v) \\ \mathcal{L}(x, v) &= \mathcal{H}(x, \cdot)^*(v) = \tilde{\mathcal{L}}(x, v + \nabla f(x)). \end{aligned} \quad (\text{A.1.17})$$

$\tilde{\mathcal{L}}$ and $\mathcal{L}$ are thus respectively equal to the Lagrangians $\mathcal{L}_Z(\cdot, \cdot)$ and $\mathcal{L}_G(\cdot, \cdot)$. Finally, (A.3.121) will be discussed in [Appendix A.3.5](#).

A.1.3 Basic properties

In this section, we derive from our assumptions some basic consequences, which will be useful throughout the proof.

We first state some properties of the Hamiltonian and the Lagrangian, which follow from their definitions.

Lemma A.1.1 (Properties of $\mathcal{H}$ and $\mathcal{L}$).

1. $\mathcal{H}$ is $\mathcal{C}^2$ and $\mathcal{H}(x, \cdot)$ is convex for any $x \in \mathbb{R}^d$.
2. $\mathcal{L}(x, \cdot)$ is convex for any $x \in \mathbb{R}^d$, $\mathcal{L}$ is l.s.c. on $\mathbb{R}^d \times \mathbb{R}^d$.
3. For any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$, $\mathcal{H}(x, v) \leq 2M(1 + \|x\|)\|v\|$ and $\text{dom } \mathcal{L}(x, \cdot) \subset \overline{\mathbb{B}}(0, 2M(1 + \|x\|))$.
4. For any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$, $\mathcal{L}(x, v) \geq 0$ and $\mathcal{L}(x, v) = 0 \iff v = \nabla f(x)$.

Proof. For the last point, since $\mathcal{H}(x, \cdot)$ is convex and differentiable,

$$\mathcal{L}(x, v) = 0 \iff 0 \in \partial_v \mathcal{L}(x, v) \iff v \in \partial_v \mathcal{H}(x, 0) \iff v = \nabla_v \mathcal{H}(x, 0), \quad (\text{A.1.18})$$

and, since the noise has zero mean, $\nabla_v \mathcal{H}(x, 0) = \nabla f(x)$. ■

Lemma A.1.2 (Growth of the iterates). *For any $x_0 \in \mathbb{R}^d$, $\eta_0 > 0$, for every $\eta \in (0, \eta_0]$, for any $N \geq 1$, $0 \leq n \leq N \lceil \eta^{-1} \rceil$,*

$$\|x_n\| \leq e^{2M(1+\eta_0)N} (1 + \|x_0\|). \quad (\text{A.1.19})$$

Proof. By the triangular inequality, for any $n \geq 0$, we have that

$$\begin{aligned} \|x_{n+1}\| &\leq \|x_n\| + \eta \|\nabla f(x_n)\| + \eta \|U_n\| \\ &\leq \|x_n\| + 2\eta M(1 + \|x_n\|) \\ &= (1 + 2\eta M)\|x_n\| + 2\eta M \end{aligned} \quad (\text{A.1.20})$$

where we used the growth condition on ∇f and U . One can then solve this recursion to show that, for any $n \geq 0$,

$$\|x_n\| \leq (1 + 2\eta M)^n (1 + \|x_0\|). \quad (\text{A.1.21})$$

The result then follows by noticing that, for $0 \leq n \leq N \lceil \eta^{-1} \rceil$,

$$\begin{aligned} (1 + 2\eta M)^n &\leq (1 + 2\eta M)^{N \lceil \eta^{-1} \rceil} \\ &\leq e^{2MN\eta \lceil \eta^{-1} \rceil} \\ &\leq e^{2MN(1+\eta)}. \quad \blacksquare \end{aligned}$$

Lemma A.1.3. *For any $x_0 \in \mathbb{R}^d$, for every $\eta \in (0, (4M)^{-1}]$, for any $N \geq 1$, $0 \leq n \leq N \lceil \eta^{-1} \rceil$,*

$$\|x_n\| \geq e^{-(4M+1)N} \|x_0\| - 1. \quad (\text{A.1.22})$$

Proof. By the triangular inequality, for any $n \geq 0$, we have that

$$\begin{aligned} \|x_{n+1}\| &\geq \|x_n\| - \eta \|\nabla f(x_n)\| - \eta \|U_n\| \\ &\leq \|x_n\| - 2\eta M(1 + \|x_n\|) \\ &= (1 - 2\eta M)\|x_n\| - 2\eta M \end{aligned} \quad (\text{A.1.23})$$

where we used the growth condition on ∇f and U . One can then solve this recursion to obtain that, for any $n \geq 0$,

$$\|x_n\| \geq (1 - 2\eta M)^n \|x_0\| - 1. \quad (\text{A.1.24})$$

The result then follows by noticing that, for $0 \leq n \leq N \lceil \eta^{-1} \rceil$,

$$\begin{aligned} (1 - 2\eta M)^n &\geq (1 - 2\eta M)^{N \lceil \eta^{-1} \rceil} \\ &\geq e^{-4MN\eta \lceil \eta^{-1} \rceil} \\ &\geq e^{-4MN(1+\eta)} \end{aligned} \quad (\text{A.1.25})$$

where we used the fact that $\log(1 - x) \geq -2x$ on $[0, 1/2]$ since $2\eta M \leq 1/2$. $\blacksquare$

A.2 A LARGE DEVIATIONS PRINCIPLE FOR SGD

A.2.1 Preliminaries

The goal of this section is to provide large deviation principles for continuous and discrete trajectories related to our algorithm of interest (A.1.15). From the

sequences $(x_n)_{n \geq 0}$ and $(\omega_n)_{n \geq 0}$, we define three sequences: one discrete (in lowercase) and two continuous (in uppercase):

(a) The discrete “rescaled” trajectory

$$x_n^\eta := x_{\lfloor n/\eta \rfloor}. \quad (\text{A.2.1a})$$

(b) The continuous “interpolated” trajectory, defined for any $n \geq 0$, $t \in [\eta n, \eta(n+1)]$ by

$$X_t = x_n + \left(\frac{t}{\eta} - n \right) (x_{n+1} - x_n) \quad (\text{A.2.1b})$$

(c) The continuous “discretized noise” trajectory, defined by $Z_0 = x_0$ and for any $t > 0$

$$\dot{Z}_t = -\nabla f(Z_t) + U(Z_t, \omega_{\lfloor t/\eta \rfloor}) \quad (\text{A.2.1c})$$

Note that all three sequences are “accelerated” by a factor $1/\eta$ compared to the original sequences appearing in (A.1.15). In this section, we establish a large deviation principle for the discrete rescaled sequence $(x_n^\eta)_{n \geq 0}$. To do so, we build upon Freidlin and Wentzell (1998b, Chap. 7) to obtain a large deviations principle on $(Z_t)_{t \geq 0}$ and then transfer it to $(X_t)_{t \geq 0}$ which enables us to obtain a discrete large deviation principle for $(x_n^\eta)_{n \geq 0}$.

We equip the space of continuous functions $\mathcal{C}_T := \mathcal{C}([0, T], \mathbb{R}^d)$ with the distance induced by the uniform norm

$$\text{dist}_{[0, T]}(\gamma, \varphi) := \sup_{t \in [0, T]} \|\gamma_t - \varphi_t\|. \quad (\text{A.2.2})$$

In order to use it as a proxy later, we will now bound the distance between the (continuous) “interpolated” trajectory and the “discretized noise” trajectory. To do so, we will first bound the latter.

Lemma A.2.1 (Growth of the trajectory). *For any $x_0 \in \mathbb{R}^d$, $\eta > 0$, $t \geq 0$, we have $\|Z_t\| \leq e^{2Mt}(\|x_0\| + 2Mt)$.*

Proof. Using the definition of Z_t in (A.2.1c) and the growth condition (A.1.16) on ∇f and U , we have that

$$\|\dot{Z}_t\| = \left\| -\nabla f(Z_t) + U(Z_t, \omega_{\lfloor t/\eta \rfloor}) \right\| \leq 2M(1 + \|Z_t\|). \quad (\text{A.2.3})$$

Hence, for any $t \geq 0$,

$$\|Z_t\| \leq \|z_0\| + \int_0^t 2M(1 + \|Z_s\|) ds = \|z_0\| + 2Mt + \int_0^t 2M\|Z_s\| ds. \quad (\text{A.2.4})$$

Invoking Grönwall’s lemma then yields the result. $\blacksquare$

The following lemma states that the distance between the “interpolated” and “discretized noise” trajectories is bounded by a factor proportional to the stepsize η .

Lemma A.2.2. *Fix $\mathcal{K} \subset \mathbb{R}^d$ compact, $\eta_0 > 0$, $T > 0$. Then, there exists some constant $c = c(\mathcal{K}, \eta_0, T, f, U, \Omega) < +\infty$ such that, for any $x_0 \in \mathcal{K}$, $\eta \in (0, \eta_0]$,*

$$\text{dist}_{[0, T]}(X, Z) \leq c \eta. \quad (\text{A.2.5})$$

Proof. Before starting, notice that [Lemma A.1.2](#) and [Lemma A.2.1](#) imply that there is some compact set $\mathcal{K}'$ that depends on $\mathcal{K}$, T , η_0 and M such that, for any $x_0 \in \mathcal{K}$, $\eta \in (0, \eta_0]$, $t \in [0, T]$, X_t and Z_t belong to $\mathcal{K}'$. In particular, U is therefore Lipschitz-continuous on $\mathcal{K}' \times \Omega$ so that, for any $\omega \in \Omega$, $x \mapsto -\nabla f(x) + U(x, \omega)$ is Lipschitz-continuous on $\mathcal{K}'$ with constant L and bounded with constant B .

Let us now estimate the derivative of the interpolated trajectory X_t , which is piecewise differentiable by definition (see [\(A.2.1b\)](#)). For any $t \in [0, T]$ such that $t \in (\eta n, \eta(n+1))$ for some $n \geq 0$, we have that

$$\begin{aligned} & \|\dot{X}_t - (-\nabla f(X_t) + U(X_t, \omega_{\lfloor t/\eta \rfloor}))\| \\ &= \left\| \frac{x_{n+1} - x_n}{\eta} - (-\nabla f(X_t) + U(X_t, \omega_{\lfloor t/\eta \rfloor})) \right\| \\ &= \|(-\nabla f(X_{\eta n}) + U(X_{\eta n}, \omega_n)) - (-\nabla f(X_t) + U(X_t, \omega_{\lfloor t/\eta \rfloor}))\| \\ &\leq L\|X_{\eta n} - X_t\| \end{aligned} \quad (\text{A.2.6})$$

where we used the L -Lipschitz-continuity of $x \mapsto -\nabla f(x) + U(x, \omega)$ on $\mathcal{K}'$ uniformly in $\omega \in \Omega$. Since this map is also bounded by B , we have that

$$\|\dot{X}_t - (-\nabla f(X_t) + U(X_t, \omega_{\lfloor t/\eta \rfloor}))\| \leq L\|X_{\eta n} - X_t\| \leq L\|x_n - x_{n+1}\| \leq LB\eta. \quad (\text{A.2.7})$$

Moreover, the L -Lipschitz-continuity of $x \mapsto -\nabla f(x) + U(x, \omega)$ also gives us that

$$\begin{aligned} & \|\dot{Z}_t - (-\nabla f(X_t) + U(X_t, \omega_{\lfloor t/\eta \rfloor}))\| \\ &= \|-\nabla f(Z_t) + U(Z_t, \omega_{\lfloor t/\eta \rfloor}) - (-\nabla f(X_t) + U(X_t, \omega_{\lfloor t/\eta \rfloor}))\| \\ &\leq L\|Z_t - X_t\|. \end{aligned} \quad (\text{A.2.8})$$

Putting everything together and integrating $X_t - Z_t$ from 0 to t yields (since X_t is absolutely continuous), for any $t \in [0, T]$,

$$\|X_t - Z_t\| \leq \|X_0 - z_0\| + \int_0^t \|\dot{X}_s - \dot{Z}_s\| ds \leq LB\eta + L \int_0^t \|X_s - Z_s\| ds \quad (\text{A.2.9})$$

where we used that $z_0 = X_0 = x_0$ by definition of the trajectories. Finally, Grönwall's lemma then yields the result. $\blacksquare$

A.2.2 Large Deviations Principle for interpolated trajectories

From the Lagrangian defined in [\(A.1.17\)](#), we define, on $\mathcal{C}_T = \mathcal{C}([0, T], \mathbb{R}^d)$, the normalized action functional $\mathcal{S}_{[0, T]}$ as

$$\mathcal{S}_{[0, T]}(\gamma) = \begin{cases} \int_0^T \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt & \text{if } \gamma \text{ is absolutely continuous,} \\ \infty & \text{otherwise,} \end{cases} \quad (\text{A.2.10})$$

following [Freidlin and Wentzell \(1998b, Chap. 3.2\)](#), as a manner to quantify how “probable” a trajectory is.

We first show that the set

$$\Gamma_{[0, T]}^{\mathcal{K}}(s) := \{\gamma \in \mathcal{C}([0, T], \mathbb{R}^d) : \gamma_0 \in \mathcal{K}, \mathcal{S}_{[0, T]}(\gamma) \leq s\} \quad (\text{A.2.11})$$

of trajectories with bounded action functional is compact and $S_{[0,T]}$ is lower semi-continuous.

Lemma A.2.3. *Fix $T > 0$. For any $\mathcal{K} \subset \mathbb{R}^d$ compact, $s \geq 0$, the set $\Gamma_{[0,T]}^{\mathcal{K}}(s)$ is compact and $S_{[0,T]}$ is l.s.c. on $\mathcal{C}([0,T], \mathbb{R}^d)$.*

Proof. Let us first check that $S_{[0,T]}$ is l.s.c. on $\mathcal{C}([0,T], \mathbb{R}^d)$ by applying [Ioffe et al. \(1979, §9.1.4, Thm. 3\)](#): $(t, x, v) \mapsto \mathcal{L}(x, v)$ is a normal integrand since $(x, v) \mapsto \mathcal{L}(x, v)$ is l.s.c. by construction, quasiregular since $\mathcal{L}(x, \cdot)$ is convex for any $x \in \mathbb{R}^d$ and satisfies the growth condition because it is non-negative (see [Lemma A.1.1](#)). Hence, $S_{[0,T]}$ is l.s.c. on $\mathcal{C}([0,T], \mathbb{R}^d)$.

The compactness of $\Gamma_{[0,T]}^{\mathcal{K}}(s)$ follows from the idea of proof as [Freidlin and Wentzell \(1998b, Chap. 7, Lemma 4.2\)](#) but with the added difficulty that the gradient ∇f and the noise U are not uniformly bounded. Take $\gamma \in \Gamma_{[0,T]}^{\mathcal{K}}(s)$. Since $S_{[0,T]}(\gamma) \leq s < +\infty$, it means that $\mathcal{L}(\gamma_t, \dot{\gamma}_t) < +\infty$ for almost every t so that by [Lemma A.1.1](#), almost everywhere,

$$\|\dot{\gamma}_t\| \leq 2M(1 + \|\gamma_t\|). \quad (\text{A.2.12})$$

Grönwall's lemma then yields that, for any $t \in [0, T]$,

$$\|\gamma_t\| \leq e^{2MT}(\|\gamma_0\| + 2MT) \quad (\text{A.2.13})$$

which is bounded uniformly in $\gamma \in \Gamma_{[0,T]}^{\mathcal{K}}(s)$ since γ_0 is in $\mathcal{K}$ compact. Hence, $\|\dot{\gamma}_t\|$ is also uniformly bounded by (A.2.12). Therefore, the functions in $\Gamma_{[0,T]}^{\mathcal{K}}(s)$ are equicontinuous and uniformly bounded and, $\Gamma_{[0,T]}^{\mathcal{K}}(s)$ is closed by l.s.c. of $S_{[0,T]}$, so that, by the Arzelà-Ascoli theorem, $\Gamma_{[0,T]}^{\mathcal{K}}(s)$ is compact. ■

The following result establishes the fact that the functional $\eta^{-1}S_{[0,T]}$ is the action functional in $\mathcal{C}([0,T], \mathbb{R}^d)$ of the interpolated process $(X_t)_{t \in [0,T]}$ of the algorithm started at x_0 , uniformly with respect to the initial point x_0 in any compact set $\mathcal{K} \subset \mathbb{R}^d$, as $\eta \rightarrow 0$. This enables to build on [Freidlin and Wentzell \(1998b, Chap. 7, Thm. 4.1'\)](#) to provide a large deviation principle for the interpolated trajectory $(X_t)_{t \in [0,T]}$, meaning that for η small enough, it will be (a) close to probable trajectories with a probability exponentially big in their action functional; and (b) far from the most probable trajectories with a probability exponentially small in their action value.

Proposition A.2.4. *Fix $T > 0$. For any $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathbb{R}^d$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, we have that*

$$\mathbb{P}_{x_0}(\text{dist}_{[0,T]}(X, \gamma) < \delta) \geq \exp\left(-\frac{S_{[0,T]}(\gamma) + \varepsilon}{\eta}\right) \quad (\text{A.2.14a})$$

$$\mathbb{P}_{x_0}(\text{dist}_{[0,T]}(X, \Gamma_{[0,T]}^{\{x_0\}}(s)) > \delta) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right) \quad (\text{A.2.14b})$$

for all $\gamma \in \Gamma_{[0,T]}^{\{x_0\}}(s)$.

Proof. Our strategy is to apply [Freidlin and Wentzell \(1998b, Chap. 7, Thm. 4.1'\)](#) to the process $(Z_t)_{t \in [0,T]}$ and use [Lemma A.2.2](#) to transfer the result to $(X_t)_{t \in [0,T]}$ (both starting from x_0). However, this theorem requires $b(x, \omega) := -\nabla f(x) + U(x, \omega)$ to be uniformly bounded as well as its derivative. To avoid this issue, note that, for a fixed compact set $\mathcal{K}$, $s > 0$, and all $\eta \leq 1$, the trajectories of

$\Gamma_{[0,T]}^{\mathcal{K}}(s)$ and of the process $(Z_t)_{t \in [0,T]}$ are contained in a compact set $\mathcal{K}'$ by the compactness of $\Gamma_{[0,T]}^{\mathcal{K}}(s)$ and [Lemma A.2.1](#).

Now, consider $b': \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}^d$ twice differentiable that coincides with b on $\mathcal{K}' \times \Omega$ but that is uniformly bounded along with its derivative. Define $\mathcal{H}'(x, v) = \log \mathbb{E}[\exp(\langle v, b'(x, \omega) \rangle)]$ which is still differentiable and satisfies “Condition F” of [Freidlin and Wentzell \(1998b, Chap. 7.4\)](#) by [Freidlin and Wentzell \(1998b, Chap. 7, Lem. 4.3\)](#) and i.i.d. assumption. Hence, [Freidlin and Wentzell \(1998b, Chap. 7, Thm. 4.1'\)](#) yields that, with $\mathcal{K}$ and s fixed above, for any $\delta, \varepsilon > 0$, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\gamma \in \Gamma_{[0,T]}^{\{x_0\}}(s)$,

$$\mathbb{P}_{x_0}(\text{dist}_{[0,T]}(Z, \gamma) < \delta) \geq \exp\left(-\frac{\mathcal{S}_{[0,T]}(\gamma) + \varepsilon}{\eta}\right) \quad (\text{A.2.15a})$$

$$\mathbb{P}_{x_0}(\text{dist}_{[0,T]}(Z, \Gamma_{[0,T]}^{\{x_0\}}(s)) > \delta) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \quad (\text{A.2.15b})$$

To obtain the result for the process $(X_t)_{t \in [0,T]}$, fix $\delta > 0$. By [Lemma A.2.2](#), there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any initial point $x_0 \in \mathcal{K}$, $\text{dist}_{[0,T]}(X, Z) < \delta$. Combining this with the previous result on $(Z_t)_{t \in [0,T]}$ yields that, for any $\delta, \varepsilon > 0$, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\gamma \in \Gamma_{[0,T]}^{\{x_0\}}(s)$,

$$\mathbb{P}_{x_0}(\text{dist}_{[0,T]}(X, \gamma) < 2\delta) \geq \exp\left(-\frac{\mathcal{S}_{[0,T]}(\gamma) + \varepsilon}{\eta}\right) \quad (\text{A.2.16a})$$

and

$$\mathbb{P}_{x_0}(\text{dist}_{[0,T]}(X, \Gamma_{[0,T]}^{\{x_0\}}(s)) > 2\delta) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right) \quad (\text{A.2.16b})$$

which concludes the proof. $\blacksquare$

A.2.3 Large Deviations Principle for discrete trajectories

We now leverage the previous section to show a LDP for the discrete rescaled trajectories $(x_n^\eta)_{n \geq 0}$ defined in [\(A.2.1a\)](#). To do so, we will use the results of the previous section by considering a finite number of points $(X_n)_{n \geq 0}$ from the continuous interpolation.

For some $N > 0$, we will first equip $(\mathbb{R}^d)^N$ with the distance

$$\text{dist}_N(\xi, \zeta) = \max_{0 \leq n \leq N-1} \|\xi_n - \zeta_n\| \quad (\text{A.2.17})$$

and bound the difference between the discrete rescaled trajectory and the continuous interpolation.

Lemma A.2.5. *Fix $\mathcal{K} \subset \mathbb{R}^d$ compact, $\eta_0 > 0$, $N \geq 1$. Then, there exists some constant $c = c(\mathcal{K}, \eta_0, N, f, U, \Omega) < +\infty$ such that, for any $x_0 \in \mathcal{K}$, $\eta \in (0, \eta_0]$,*

$$\text{dist}_N(x^\eta, (X_n)_{0 \leq n \leq N-1}) = \max_{0 \leq n \leq N-1} \|x_n^\eta - X_n\| \leq c\eta. \quad (\text{A.2.18})$$

Proof. By [Lemma A.1.2](#), there is some compact set $\mathcal{K}'$ that depends on $\mathcal{K}$, N , η_0 and M such that, for any $x_0 \in \mathcal{K}$, $\eta \in (0, \eta_0]$, $n \leq \lceil 1/\eta \rceil N$, x_n belongs to $\mathcal{K}'$. In particular, for almost every $t \in [0, N - 1]$, we have, that,

$$\dot{X}_t = \frac{x_{\lfloor t/\eta \rfloor + 1} - x_{\lfloor t/\eta \rfloor}}{\eta} = -\nabla f(x_{\lfloor t/\eta \rfloor}) + U(x_{\lfloor t/\eta \rfloor}, \omega_{\lfloor t/\eta \rfloor}) \quad (\text{A.2.19})$$

with $x_{\lfloor t/\eta \rfloor}$ belonging to $\mathcal{K}'$ since $\left\lfloor \frac{t}{\eta} \right\rfloor \leq \left\lfloor \frac{N-1}{\eta} \right\rfloor \leq \left\lceil \frac{N}{\eta} \right\rceil$. Hence, the norm of $\dot{X}_t$ is bounded by some constant B for almost every $t \in [0, N - 1]$ uniformly in $x_0 \in \mathcal{K}$, $\eta \in (0, \eta_0]$ by the growth condition in [\(b\)](#).

Therefore, for any $0 \leq n \leq N - 1$, since $x_n^\eta = x_{n\lfloor 1/\eta \rfloor} = X_{n\lfloor 1/\eta \rfloor}$, we have that

$$\|x_n^\eta - X_n\| \leq Bn|1 - \eta\lfloor 1/\eta \rfloor| \leq BN\eta \quad (\text{A.2.20})$$

which concludes the proof. $\blacksquare$

Now, for $N \geq 0$, $\xi = (\xi_0, \dots, \xi_{N-1}) \in (\mathbb{R}^d)^N$, let us define the normalized discrete action functional

$$\mathcal{S}_N(\xi) := \sum_{n=0}^{N-2} \rho(\xi_n, \xi_{n+1}) \quad (\text{A.2.21})$$

where the cost of moving from one iteration to the next is defined for any $x, x' \in \mathbb{R}^d$ from the previous continuous normalized action functional (cf. [\(A.2.10\)](#)) with horizon 1 as

$$\rho(x, x') := \inf\{\mathcal{S}_{0,1}(\gamma) : \gamma \in \mathcal{C}([0, 1], \mathbb{R}^d), \gamma_0 = x, \gamma_1 = x'\}. \quad (\text{A.2.22})$$

Again, we show that the set of discrete trajectories with low action functional

$$\Gamma_N^{\mathcal{K}}(s) := \{\xi \in (\mathbb{R}^d)^N : \xi_0 \in \mathcal{K}, \mathcal{S}_N(\xi) \leq s\} \quad (\text{A.2.23})$$

is compact and $\mathcal{S}_N$ is l.s.c..

Lemma A.2.6. Fix $N \geq 0$. For any $\mathcal{K} \subset \mathbb{R}^d$ compact, $s \geq 0$, the set $\Gamma_N^{\mathcal{K}}(s)$ is compact and $\mathcal{S}_N$ is l.s.c. on $(\mathbb{R}^d)^N$.

Proof. First, let us show that, for $s \geq 0$, $\mathcal{K} \subset \mathbb{R}^d$, the set $\{(x, x') \in \mathcal{K} \times \mathbb{R}^d : \rho(x, x') \leq s\}$ is compact. Let $(x^k, x'^k)_{k \geq 0}$ be a sequence in $\mathcal{K} \times \mathbb{R}^d$ such that $\rho(x^k, x'^k) \leq s$ for all $k \geq 0$. Then, for any $k \geq 0$, there exists $\gamma^k \in \mathcal{C}([0, 1], \mathbb{R}^d)$ such that $\gamma_0^k = x^k$, $\gamma_1^k = x'^k$ and $\mathcal{S}_{0,1}(\gamma^k) \leq \rho(x^k, x'^k) + (1 + k)^{-1}$. In particular, for any $k \geq 0$, γ^k belongs to $\{\gamma \in \mathcal{C}([0, 1], \mathbb{R}^d) : \gamma_0 \in \mathcal{K}, \mathcal{S}_{0,1}(\gamma) \leq s + 1\}$ which is compact by [Lemma A.2.3](#). Hence, there exists a subsequence that converges uniformly to some $\gamma \in \mathcal{C}([0, 1], \mathbb{R}^d)$. Without loss of generality, assume that $\gamma^k \rightarrow \gamma$ as $k \rightarrow \infty$. In particular, (x^k, x'^k) converges to (γ_0, γ_1) as $k \rightarrow \infty$ with γ_0 belonging to $\mathcal{K}$. Then, by l.s.c. of $\mathcal{S}_{0,1}$, we have that $\mathcal{S}_{0,1}(\gamma) \leq \liminf_{k \rightarrow \infty} \rho(x^k, x'^k) \leq s$ so that $\rho(\gamma_0, \gamma_1) \leq s$. Therefore, we have shown that $\{(x, x') \in \mathcal{K} \times \mathbb{R}^d : \rho(x, x') \leq s\}$ is compact.

As a consequence ρ is l.s.c. on $\mathbb{R}^d \times \mathbb{R}^d$: indeed, for any $s \geq 0$, any convergent sequence of points of $\{(x, x') \in \mathbb{R}^d \times \mathbb{R}^d : \rho(x, x') \leq s\}$ must be included in $\{(x, x') \in \mathcal{K} \times \mathbb{R}^d : \rho(x, x') \leq s\}$ for some $\mathcal{K} \subset \mathbb{R}^d$ compact, which is closed and therefore contains the limit point of any such sequences. $\mathcal{S}_N$ is then immediately l.s.c. on $(\mathbb{R}^d)^N$.

Finally, the compactness of $\Gamma_N^{\mathcal{K}}(s)$ follows by induction on N . For $N = 1$, $\Gamma_1^{\mathcal{K}}(s) = \{(x, x') \in \mathbb{R}^d \times \mathbb{R}^d : \rho(x, x') \leq s\}$ which is compact by the previous

argument for any compact set $\mathcal{K}$. Now, assume that $\Gamma_N^\mathcal{K}(s)$ is compact for some $N \geq 1$. As a consequence, the set

$$\mathcal{K}' := \{x \in \mathbb{R}^d : \exists (\xi_0, \dots, \xi_{N-2}) \in (\mathbb{R}^d)^{N-1}, (\xi_0, \dots, \xi_{N-2}, x) \in \Gamma_N^\mathcal{K}(s)\} \quad (\text{A.2.24})$$

is compact as well. Hence, $\Gamma_{N+1}^\mathcal{K}(s)$ is included in the product of compact sets $\Gamma_{N-1}^\mathcal{K}(s) \times \Gamma_1^{\mathcal{K}'}(s)$ and is therefore bounded. Moreover, $\Gamma_{N+1}^\mathcal{K}(s)$ is closed by l.s.c. of $\mathcal{S}_N$ and therefore compact. This concludes the proof by induction and the proof of the lemma. $\blacksquare$

We will first provide a discrete analogue of [Proposition A.2.4](#) for the interpolated trajectory at times $n = 0, \dots, N-1$, with $\eta^{-1}\mathcal{S}_N$ the action functional in $(\mathbb{R}^d)^N$ of the process $(X_n)_{0 \leq n \leq N-1}$, uniformly with respect to the starting point x_0 in any compact set $\mathcal{K} \subset \mathbb{R}^d$, as $\eta \rightarrow 0$.

Proposition A.2.7. *Fix $N \geq 0$. For any $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathbb{R}^d$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\xi \in \Gamma_N^{\{x_0\}}(s)$, we have that*

$$\mathbb{P}_{x_0}(\text{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{S}_N(\xi) + \varepsilon}{\eta}\right) \quad (\text{A.2.25a})$$

and

$$\mathbb{P}_{x_0}(\text{dist}_N((X_n)_{0 \leq n \leq N-1}, \Gamma_N^{\{x_0\}}(s)) > \delta) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \quad (\text{A.2.25b})$$

for all $\xi \in \Gamma_N^{\{x_0\}}(s)$

Proof. Invoke the first part of [Proposition A.2.4](#) with $T \leftarrow N-1$ and $s \leftarrow s + \varepsilon$. There exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\gamma \in \Gamma_{0,N}^{\{x_0\}}(s + \varepsilon)$,

$$\mathbb{P}_{x_0}(\text{dist}_{0,N-1}(X, \gamma) < \delta) \geq \exp\left(-\frac{\mathcal{S}_{0,N-1}(\gamma) + \varepsilon}{\eta}\right). \quad (\text{A.2.26})$$

Take $\eta \in (0, \eta_0]$, $x_0 \in \mathcal{K}$, $\xi \in \Gamma_N^{\{x_0\}}(s)$. Then, there exists $\gamma \in \Gamma_{0,N}^{\{x_0\}}(s + \varepsilon)$, for any $0 \leq n \leq N-1$, $\gamma_n = \xi_n$, cf. [\(A.2.22\)](#). Hence,

$$\begin{aligned} \mathbb{P}_{x_0}(\text{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) < \delta) &\geq \mathbb{P}_{x_0}(\text{dist}_{0,N-1}(X, \gamma) < \delta) \\ &\geq \exp\left(-\frac{\mathcal{S}_{0,N-1}(\gamma) + \varepsilon}{\eta}\right) \\ &\geq \exp\left(-\frac{\mathcal{S}_N(\xi) + 2\varepsilon}{\eta}\right) \end{aligned} \quad (\text{A.2.27})$$

which prove the first part of the result.

For the second part, we have similarly from [Proposition A.2.4](#) with $T \leftarrow N-1$ and $\delta \leftarrow \delta/2$ that for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$,

$$\mathbb{P}_{x_0}(\text{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s)) > \delta/2) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \quad (\text{A.2.28})$$

Now, note that if $\text{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s)) < \delta$, then there must exist $\gamma \in \Gamma_{0,N-1}^{\{x_0\}}(s)$ such that $\text{dist}_{0,N-1}(X, \gamma) \leq \delta$. Consider the discrete path $\xi \in (\mathbb{R}^d)^N$ defined by

$\xi_n = \gamma_n$ for $0 \leq n \leq N-1$. Then, by construction, $\mathcal{S}_N(\xi) \leq \mathcal{S}_{0,N-1}(\gamma) \leq s$ and $\text{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) \leq \delta$. Thus,

$$\text{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s)) < \delta \implies \text{dist}_N((X_n)_{0 \leq n \leq N-1}, \Gamma_N^{\{x_0\}}(s)) \leq \delta. \quad (\text{A.2.29})$$

Putting all together, we have that

$$\begin{aligned} \mathbb{P}_{x_0}(\text{dist}_N((X_n)_{0 \leq n \leq N-1}, \Gamma_N^{\{x_0\}}(s)) > \delta) &\leq \mathbb{P}_{x_0}(\text{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s) \geq \delta) \\ &\leq \mathbb{P}_{x_0}(\text{dist}_{0,N-1}(X, \Gamma_{0,N-1}^{\{x_0\}}(s) > \delta/2) \\ &\leq \exp\left(-\frac{s-\varepsilon}{\eta}\right) \end{aligned} \quad (\text{A.2.30})$$

which concludes our proof. $\blacksquare$

Finally, we end up with a large deviation principle on the discrete rescaled iterates $(x_n^\eta)_{0 \leq n \leq N-1} = (x_{n \lfloor \eta^{-1} \rfloor})_{0 \leq n}$ by leveraging [Lemma A.2.5](#). In the following result, the functional $\eta^{-1}\mathcal{S}_N$ is thus the action functional in $(\mathbb{R}^d)^N$ of the process $(x_n^\eta)_{0 \leq n \leq N-1}$ uniformly with respect to the starting point x_0 in any compact set $\mathcal{K} \subset \mathbb{R}^d$, as $\eta \rightarrow 0$.

Corollary A.2.8. *Fix $N \geq 0$. For any $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathbb{R}^d$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\xi \in \Gamma_N^{\{x_0\}}(s)$, we have that*

$$\mathbb{P}_{x_0}(\text{dist}_N(x^\eta, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{S}_N(\xi) + \varepsilon}{\eta}\right) \quad (\text{A.2.31a})$$

and

$$\mathbb{P}_{x_0}(\text{dist}_N(x^\eta, \Gamma_N^{\{x_0\}}(s)) > \delta) \leq \exp\left(-\frac{s-\varepsilon}{\eta}\right). \quad (\text{A.2.31b})$$

for all $\xi \in \Gamma_N^{\{x_0\}}(s)$.

Proof. Fix $\delta > 0$. Choose η_0 such that both [Proposition A.2.7](#) and [Lemma A.2.5](#) hold with $c\eta \leq \delta$. Then for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $\xi \in \Gamma_N^{\{x_0\}}(s)$, $\xi \in \Gamma_N^{\{x_0\}}(s)$, we have that

$$\mathbb{P}_{x_0}(\text{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{S}_N(\xi) + \varepsilon}{\eta}\right). \quad (\text{A.2.32})$$

Now, if $\text{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) < \delta$ and $\text{dist}_N(x^\eta, (X_n)_{0 \leq n \leq N-1}) \leq \delta$, then $\text{dist}_N(x^\eta, \xi) < 2\delta$. Thus,

$$\mathbb{P}_{x_0}(\text{dist}_N(x^\eta, \xi) < 2\delta) \geq \mathbb{P}_{x_0}(\text{dist}_N((X_n)_{0 \leq n \leq N-1}, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{S}_N(\xi) + \varepsilon}{\eta}\right). \quad (\text{A.2.33})$$

The second part can be obtained similarly. $\blacksquare$

To summarize this part on large deviations principles, we state a corollary containing all the results that will be needed in the following sections.

Corollary A.2.9. *Fix $N \geq 0$. Then:*

- For all $s > 0$, the set

$$\Gamma_N^{\mathcal{K}}(s) := \{\xi \in (\mathbb{R}^d)^N : \xi_0 \in \mathcal{K}, \mathcal{S}_N(\xi) \leq s\} \quad (\text{A.2.34})$$

is compact and $\mathcal{S}_N$ is l.s.c. on $(\mathbb{R}^d)^N$.

- For all $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathbb{R}^d$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for all $x_0 \in \mathcal{K}$, $n \leq N$, $\xi \in \Gamma_n^{\{x_0\}}(s)$, we have that

$$\mathbb{P}_{x_0}(\text{dist}_n(x^\eta, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{S}_n(\xi) + \varepsilon}{\eta}\right) \quad (\text{A.2.35a})$$

and

$$\mathbb{P}_{x_0}(\text{dist}_n(x^\eta, \Gamma_n^{\{x_0\}}(s)) > \delta) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \quad (\text{A.2.35b})$$

A.3 ATTRACTORS AND LIMITING MEASURES VIA LARGE DEVIATIONS

We now take inspiration from the framework of [Kifer \(1988a\)](#) in order to relate the sets of critical points to the sets where points can move at no cost. Then, we relate the probability of SGD moving to neighborhoods of critical sets to the probability of being close to well-chosen paths, which enables us to use the results of the previous section. Finally, we build upon these results to provide bounds on the limiting measure of SGD.

A.3.1 Setup

We first need to define the gradient flow of f .

Definition A.1. Define, for $x \in \mathbb{R}^d$, the flow Θ of $-\nabla f$ starting at x , i.e.,

$$\dot{\Theta}_t(x) = -\nabla f(\Theta_t(x)) \quad \text{with} \quad \Theta_0(x) = x \quad (\text{A.3.1})$$

and let $F(x)$ be the value of this flow at time 1, i.e.,

$$F(x) = \Theta_1(x). \quad (\text{A.3.2})$$

Lemma A.3.1 (Properties of the flow). Θ is well-defined and continuous in both time and space, and, for any $T \geq 0$, $\gamma \in \mathcal{C}([0, T], \mathbb{R}^d)$ such that $\gamma_0 = x$,

$$\mathcal{S}_{0,T}(\gamma) = 0 \iff \gamma_t = \Theta_t(x) \text{ for all } t \in [0, T]. \quad (\text{A.3.3})$$

Proof. The well-definition and continuity of Θ are a consequence of f being twice continuously differentiable and of the global Cauchy-Lipschitz (Picard–Lindelöf) theorem for ordinary differential equation (ODE). The equivalence follows from the uniqueness of the flow and [Lemma A.1.1](#) since

$$\begin{aligned} \mathcal{S}_{0,T}(\gamma) = 0 &\iff \mathcal{L}(\gamma_t, \dot{\gamma}_t) = 0 \text{ almost everywhere} \\ &\iff \dot{\gamma}_t = -\nabla f(\gamma_t) \text{ almost everywhere} \end{aligned} \quad (\text{A.3.4})$$

and thus, by extending $\dot{\gamma}$ by continuity, both γ and Θ satisfy the same ODE with the same initial condition and are thus equal for all t by uniqueness of the solutions. $\blacksquare$

The following lemma translates this for F .

Lemma A.3.2 (Properties of F). F is well-defined and continuous and, for any $x, x' \in \mathbb{R}^d$,

$$\rho(x, x') = 0 \iff x' = F(x). \quad (\text{A.3.5})$$

Proof. The implication ($\Leftarrow$) is immediate by definition of F and [Lemma A.3.1](#). Now for the reverse, assume that $\rho(x, x') = 0$. Following the proof of [Lemma A.2.6](#), there exists $\gamma \in \mathcal{C}([0, 1], \mathbb{R}^d)$ such that $\gamma_0 = x$, $\gamma_1 = x'$ and $\mathcal{S}_{0,1}(\gamma) = 0$. By [Lemma A.3.1](#), $\gamma_t = \Theta_t(x)$ for all $t \in [0, 1]$ and thus $x' = \Theta_1(x) = F(x)$. ■

A.3.2 Attractors

Let us first formalize the minimum-energy displacement between two points.

Definition A.2 ([Kifer \(1988a, §1.5\)](#)). Define, for $x, x' \in \mathbb{R}^d$,

$$\begin{aligned} \mathcal{B}(x, x') &= \inf \{ \mathcal{S}_{[0,T]}(\gamma) : \gamma \in \mathcal{C}([0, T], \mathbb{R}^d), \gamma_0 = x, \gamma_T = x', T \in \mathbb{N}, T \geq 1 \} \\ &= \inf \{ \mathcal{S}_N(\xi) : \xi \in (\mathbb{R}^d)^N, \xi_0 = x, \xi_{N-1} = x', N \geq 1 \}. \end{aligned} \quad (\text{A.3.6})$$

The fact that these two expressions coincide directly come from the definition of ρ .

This enables us to define an equivalence relation for the critical points of f by grouping points connected by a null-energy path.

Proposition A.3.3. *The relation $\sim$ defined for any $x, x' \in \text{crit } f$ as*

$$x \sim x' \iff \mathcal{B}(x, x') = \mathcal{B}(x', x) = 0 \quad (\text{A.3.7})$$

is an equivalence relation on $\text{crit } f$.

Proof. • Reflexivity: $\mathcal{B}(x, x) = 0$ by [Lemma A.3.1](#) since the flow started at $x \in \text{crit } f$ is constant.

- Symmetry: this follows from the definition of $\sim$.
- Transitivity: for any $x, x', x'' \in \text{crit } f$, we have by construction of $\mathcal{B}$ that

$$\mathcal{B}(x, x'') \leq \mathcal{B}(x, x') + \mathcal{B}(x', x''). \quad (\text{A.3.8})$$

Therefore, if $x \sim x'$ and $x' \sim x''$, then $\mathcal{B}(x, x'') = 0$. $\mathcal{B}(x'', x) = 0$ follows with a symmetric argument and thus $x \sim x''$. ■

Near critical points of f , the Lagrangian $\bar{\mathcal{L}}$ is actually very regular.

Lemma A.3.4. *For any $x \in \mathbb{R}^d$, there exists $\delta > 0$ such that $\bar{\mathcal{L}}$ is finite and jointly Lipschitz continuous on $\mathbb{B}(x, \delta) \times \mathbb{B}(0, \delta)$.*

Moreover, the following supremum is finite:

$$\sup \left\{ \frac{\mathcal{L}(x', v)}{\|v\|^2} : x' \in \mathbb{B}(x, \delta) \cap \text{crit } f, v \in \mathbb{B}(0, \delta) \right\} < \infty. \quad (\text{A.3.9})$$

Proof. Take $x \in \mathbb{R}^d$. We apply the implicit function theorem to the equation

$$\nabla_v \bar{\mathcal{H}}(x', w) = v, \quad (\text{A.3.10})$$

in the variables $(x', v, w) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$.

We derive that

$$\nabla_v \bar{\mathcal{H}}(x', v) = \frac{\mathbb{E}[U(x', \omega) \exp(\langle v, U(x', \omega) \rangle)]}{\mathbb{E}[\exp(\langle v, U(x', \omega) \rangle)]} \quad (\text{A.3.11})$$

and thus $(x, 0, 0)$ is solution of (A.3.10) since

$$\nabla_v \bar{\mathcal{H}}(x, 0) = \mathbb{E}[U(x, \omega)] = 0. \quad (\text{A.3.12})$$

Moreover, $\text{Hess}_v \bar{\mathcal{H}}(x, 0) = \mathbb{E}[U(x, \omega)U(x, \omega)^\top]$ which is positive definite and thus invertible by the blanket assumptions.

Hence, we can apply the implicit function theorem to get that there exists $\delta > 0$, $w: \mathbb{B}(x, \delta) \times \mathbb{B}(0, \delta) \rightarrow \mathbb{R}^d \mathcal{C}^2$ such that, for any $x' \in \mathbb{B}(x, \delta)$, $v \in \mathbb{B}(0, \delta)$,

$$\nabla_v \bar{\mathcal{H}}(x', w(x', v)) = v. \quad (\text{A.3.13})$$

Therefore, for any $x' \in \mathbb{B}(x, \delta)$, $v \in \mathbb{B}(0, \delta)$, since $\bar{\mathcal{L}}(x', v) = \bar{\mathcal{H}}(x', \cdot)^*(v)$, we have

$$\bar{\mathcal{L}}(x', v) = \langle v, w(x', v) \rangle - \bar{\mathcal{H}}(x', w(x', v)), \quad (\text{A.3.14})$$

which is finite and $\mathcal{C}^2$ on $\mathbb{B}(x, \delta) \times \mathbb{B}(0, \delta)$. Therefore, $\bar{\mathcal{L}}$ is actually L -jointly Lipschitz continuous on $\bar{\mathbb{B}}(x, \delta/2) \times \bar{\mathbb{B}}(0, \delta/2)$.

For the second part of the lemma, note that the implicit function theorem also ensures that there is $\mathcal{C}^\delta \subset \mathbb{R}^d$ a neighborhood of 0 such that, for any $x' \in \mathbb{B}(x, \delta)$, $v \in \mathbb{B}(0, \delta)$, $w(x', v)$ is the unique solution of (A.3.10) in $\mathcal{C}^\delta$. But, for any $x' \in \mathbb{B}(x, \delta)$ and $v = 0$, $w = 0$ is a solution of (A.3.10) in $\mathcal{C}^\delta$ so that necessarily $w(x', 0) = 0$, and, as a consequence, $\nabla_v \bar{\mathcal{L}}(x', 0) = 0$.

Hence, for any $x' \in \bar{\mathbb{B}}(x, \delta/2)$, $v \in \bar{\mathbb{B}}(0, \delta/2)$,

$$\begin{aligned} \bar{\mathcal{L}}(x', v) &= \bar{\mathcal{L}}(x', v) - \bar{\mathcal{L}}(x', 0) - \langle v, \nabla_v \bar{\mathcal{L}}(x', 0) \rangle \\ &\leq \frac{1}{2} \sup_{\bar{\mathbb{B}}(x, \delta/2) \times \bar{\mathbb{B}}(0, \delta/2)} \|\text{Hess}_v \bar{\mathcal{L}}\| \|v\|^2. \end{aligned} \quad (\text{A.3.15})$$

To conclude, it suffices to note that, for any $x' \in \mathbb{B}(x, \delta) \cap \text{crit } f$, $\nabla f(x) = 0$ and therefore $\mathcal{L}(x', \cdot)$ and $\bar{\mathcal{L}}(x', \cdot)$ coincide. $\blacksquare$

Lemma A.3.5. *For any $x \in \text{crit}(f)$, for any $\epsilon > 0$, there exists $\delta > 0$ such that, for any $x' \in \mathbb{B}(x, \delta)$, $\rho(x, x') \leq \epsilon$ and $\rho(x', x) \leq \epsilon$.*

Proof. Take $x \in \mathbb{R}^d$ such that $\nabla f(x) = 0$.

By Lemma A.3.4, there exists $\delta > 0$ such that $\bar{\mathcal{L}}$ is finite and L -jointly Lipschitz continuous on $\mathbb{B}(x, \delta) \times \mathbb{B}(0, \delta)$. In particular, for any $x' \in \mathbb{B}(x, \delta)$, $v \in \mathbb{B}(0, \delta)$,

$$\bar{\mathcal{L}}(x', v) \leq L\|v\|. \quad (\text{A.3.16})$$

For any integer $k \geq 2$, by continuity of ∇f (Assumption A.1), there is $0 < \delta'_k < \delta/k$ such that, for every $x' \in \mathbb{B}(x, \delta'_k)$, $\|\nabla f(x')\| \leq \delta/k$. Then, for any $x' \in \mathbb{B}(x, \delta'_k)$, $v \in \mathbb{B}(0, \delta'_k)$,

$$\begin{aligned} \mathcal{L}(x', v) &= \bar{\mathcal{L}}(x', v + \nabla f(x')) \\ &\leq L(\|v + \nabla f(x')\|) \\ &\leq L(\|v\| + \|\nabla f(x')\|) \\ &\leq \frac{2L\delta}{k}. \end{aligned} \quad (\text{A.3.17})$$

Take $x' \in \mathbb{B}(x, \delta'_k)$ and define the path $\gamma \in \mathcal{C}([0, 1], \mathbb{R}^d)$ by $\gamma_t = x + t(x' - x)$. Since for any $t \in [0, 1]$, $\gamma_t \in \mathbb{B}(x, \delta'_k)$ and $\dot{\gamma}_t = x' - x \in \mathbb{B}(0, \delta'_k)$, we can bound the cost as

$$\rho(x, x') \leq \mathcal{S}_{0,1}(\gamma) = \int_0^1 \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt \leq \frac{2L\delta}{k}. \quad (\text{A.3.18})$$

As a result, taking k large enough yields the existence of $\delta'_k > 0$ such that, for any $x' \in \mathbb{B}(x, \delta'_k)$, $\rho(x, x') \leq \epsilon$. The reverse inequality is obtained by the same argument. ■

The following lemma relates the Lagrangian to the gradient and our noise structure.

Lemma A.3.6. *For any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$,*

$$\mathcal{L}(x, v) \geq \frac{\|v + \nabla f(x)\|^2}{2\sigma_\infty^2(f(x))}. \quad (\text{A.3.19})$$

Proof. For any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$, we have that

$$\mathcal{H}(x, v) = -\langle v, \nabla f(x) \rangle + \bar{\mathcal{H}}(x, v) \leq -\langle v, \nabla f(x) \rangle + \frac{1}{2}\sigma_\infty^2(f(x))\|v\|^2. \quad (\text{A.3.20})$$

Taking the conjugate then yields that

$$\mathcal{L}(x, v) \geq \frac{\|v + \nabla f(x)\|^2}{2\sigma_\infty^2(f(x))}. \quad (\text{A.3.21})$$

■

Now, let us define a potential function V_∞ on $\mathbb{R}^d$ that uses the minimal displacement energy between two points that will be heavily used in the proofs.

Definition A.3 (Potential). Define, for $x \in \mathbb{R}^d$

$$V_\infty(x) = 2\alpha_\infty(f(x)) \quad (\text{A.3.22})$$

where $\alpha_\infty : \mathbb{R} \rightarrow \mathbb{R}$ is a twice continuously differentiable primitive of $1/\sigma_\infty^2$.

Lemma A.3.7. *For any $x, x' \in \mathbb{R}^d$,*

$$V_\infty(x') - V_\infty(x) \leq \mathcal{B}(x, x'). \quad (\text{A.3.23})$$

Proof. By [Definition A.2](#), there exists $T \geq 1$, $\gamma \in \mathcal{C}([0, T], \mathbb{R}^d)$ such that $\gamma_0 = x$, $\gamma_T = x'$ and $\mathcal{S}_{[0,T]}(\gamma) \leq \mathcal{B}(x, x') + \epsilon$. Then, by Young's inequality, we have that

$$\begin{aligned} V_\infty(x') - V_\infty(x) &= 2 \int_0^T \frac{\langle \dot{\gamma}_t, \nabla f(\gamma_t) \rangle}{\sigma_\infty^2(f(\gamma_t))} dt \\ &\leq \int_0^T \left(\frac{\langle \dot{\gamma}_t, \nabla f(\gamma_t) \rangle}{\sigma_\infty^2(f(\gamma_t))} + \frac{\|\nabla f(\gamma_t)\|^2 + \|\dot{\gamma}_t\|^2}{2\sigma_\infty^2(f(\gamma_t))} \right) dt \\ &= \int_0^T \frac{\|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2}{2\sigma_\infty^2(f(\gamma_t))} dt \\ &\leq \int_0^T \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt = \mathcal{S}_{[0,T]}(\gamma) \end{aligned} \quad (\text{A.3.24})$$

where we used [Lemma A.3.6](#) in the last inequality. Finally, our choice of γ implies that

$$V_\infty(x') - V_\infty(x) \leq \mathcal{B}(x, x') + \varepsilon \quad (\text{A.3.25})$$

which concludes the proof. $\blacksquare$

Lemma A.3.8 (Equivalence classes are closed). *Equivalence classes of $\sim$ are closed in $\mathbb{R}^d$. As a consequence, equivalence classes are compact.*

Proof. Let $x \in \text{crit } f$, and take any sequence $(x'_k)_{k \geq 0}$ in $\text{crit } f$ such that $x \sim x'_k$ for every $k \geq 0$ and which converges to some $x' \in \mathbb{R}^d$. To show that the equivalence classes are closed, we need to show that $x \sim x'$; then compactity follows directly since the equivalence classes are subsets of $\text{crit } f$ which is compact by assumption.

Since $\text{crit } f$ is closed, it holds that x' belongs to $\text{crit } f$. We now show that both $\mathcal{B}(x, x')$ and $\mathcal{B}(x', x)$ are null. We only show that $\mathcal{B}(x, x') = 0$ since the proof for the other equality is symmetric.

By [Lemma A.3.5](#), for any $\varepsilon > 0$ there is a neighborhood of x' on which $\rho(\cdot, x') \leq \varepsilon$. Take k large enough so that x'_k belongs to this neighborhood. Since $x \sim x'_k$, there exists $N \geq 1$, $\xi \in (\mathbb{R}^d)^N$ such that $\xi_0 = x$, $\xi_{N-1} = x'_k$ and $S_N(\xi) \leq \varepsilon$. Then, the path $\zeta := (\xi_0, \dots, \xi_{N-1}, x') \in (\mathbb{R}^d)^{N+1}$ and satisfies $\zeta_0 = x$, $\zeta_N = x'$ and

$$S_{N+1}(\zeta) = S_N(\xi) + \rho(x'_k, x') \leq 2\varepsilon. \quad (\text{A.3.26})$$

Hence, we have shown that, for any $\varepsilon > 0$, $\mathcal{B}(x, x') \leq 2\varepsilon$ so that $\mathcal{B}(x, x') = 0$. $\blacksquare$

Lemma A.3.9. *For any set $\mathcal{C} \subset \text{crit } f$ and $x \in \mathbb{R}^d$, define*

$$\begin{aligned} \rho(x, \mathcal{C}) &:= \inf_{x' \in \mathcal{C}} \rho(x, x') \\ \rho(\mathcal{C}, x) &:= \inf_{x' \in \mathcal{C}} \rho(x', x). \end{aligned} \quad (\text{A.3.27})$$

Then, for any $\mathcal{C} \subset \text{crit } f$ and $r > 0$, the set

$$\mathcal{W}_r(\mathcal{C}) := \{x \in \mathbb{R}^d : \rho(x, \mathcal{C}) < r, \rho(\mathcal{C}, x) < r\} \quad (\text{A.3.28})$$

is a neighborhood of $\mathcal{C}$.

Proof. [Lemma A.3.5](#) ensures that, for any $x \in \mathcal{C}$, there exists $\delta_x > 0$ such that, for any $x' \in \mathbb{B}(x, \delta_x)$, $\rho(x, x') < r$ and $\rho(x', x) < r$. $\mathcal{C}^\delta := \bigcup_{x \in \mathcal{C}} \mathbb{B}(x, \delta_x)$ is then an open set that contains $\mathcal{C}$ and is included in $\mathcal{W}_r(\mathcal{C})$. $\blacksquare$

This lemma is adapted and significantly expanded from [Kifer \(1988a, §1.5, Lem. 5.2\)](#) to handle both the unboundedness of the space and the fact that $\mathcal{B}$ is neither l.s.c. nor u.s.c..

Lemma A.3.10. *Let $\mathcal{K}$ be an equivalence class of $\sim$. Then, for any $\varepsilon > 0$, there is some $N \geq 1$ such that, for any $x, z \in \mathcal{K}$, there is $\xi \in (\mathbb{R}^d)^N$ such that $\xi_0 = x$, $\xi_{N-1} = z$, $S_N(\xi) < \varepsilon$ and $\max_{0 \leq n < N} d(\xi_n, \mathcal{K}) < \varepsilon$.*

Proof. Fix $\varepsilon > 0$. $\mathcal{K}$ is made of critical points of f so that by [Lemma A.3.5](#), for any $x \in \mathcal{K}$, there exists $\delta_x > 0$ such that, for any $x' \in \mathbb{B}(x, \delta_x)$, $\rho(x, x') < \varepsilon$ and $\rho(x', x) < \varepsilon$.

Since the open sets $\mathbb{B}(x, \delta_x)$ for $x \in \mathcal{K}$ form an open cover of $\mathcal{K}$ and $\mathcal{K}$ is compact ([Lemma A.3.8](#)), there exists a finite number of points $x_i \in \mathcal{K}$, $i \in I$ such that

$$\mathcal{K} \subset \bigcup_{i \in I} \mathbb{B}(x_i, \delta_{x_i}). \quad (\text{A.3.29})$$

We first show that the result holds for the points x_i before explaining why it actually suffices for the general case.

Fix $i, j \in I$. Since $x_i \sim x_j$, $\mathcal{B}(x_i, x_j) = 0$ and therefore there exists sequences $\xi^k \in (\mathbb{R}^d)^{N_k}$ with $\xi_0^k = x_i$, $\xi_{N_k-1}^k = x_j$ such that $\mathcal{S}_{N_k}(\xi^k) \rightarrow 0$ as $k \rightarrow \infty$. For the sake of contradiction, assume that from some k , there always exists $0 \leq n < N_k$ such that $d(\xi_n^k, \mathcal{K}) \geq \varepsilon$. Define n_k as the smallest n such that it happens, which is necessarily greater or equal to 1. Note that, by definition, $\xi_{n_k-1}^k$ satisfies $d(\xi_{n_k-1}^k, \mathcal{K}) \leq \varepsilon$.

Define $\mathcal{K}' := \{x \in \mathbb{R}^d : d(x, \mathcal{K}) \leq \varepsilon\}$ which is compact. However, for k large enough, $\rho(\xi_{n_k-1}^k, \xi_{n_k}^k) \leq \mathcal{S}_{N_k}(\xi^k) \leq 1$ so that $(\xi_{n_k-1}^k, \xi_{n_k}^k)$ belongs to $\Gamma_2^{\mathcal{K}'}(1)$, which is compact by [Corollary A.2.9](#). Therefore, one can extract a subsequence from $(\xi_{n_k-1}^k, \xi_{n_k}^k)_{k \geq 1}$ that converges to some $(x, z) \in (\mathbb{R}^d)^2$ that satisfies $d(z, \mathcal{K}) \geq \varepsilon$. Without loss of generality, at the cost of replacing the original sequence by the extracted subsequence, we assume that the sequence $(\xi_{n_k-1}^k, \xi_{n_k}^k)_{k \geq 1}$ itself converges to (x, z) . By l.s.c. of ρ , one has that

$$\begin{aligned} \rho(x, z) &\leq \liminf_{k \rightarrow \infty} \rho(\xi_{n_k-1}^k, \xi_{n_k}^k) \\ &\leq \liminf_{k \rightarrow \infty} \mathcal{S}_{N_k}(\xi^k) = 0, \end{aligned} \quad (\text{A.3.30})$$

so that $\rho(x, z) = 0$. We now show that $x = z$ and that it is a critical point. By [Lemma A.3.7](#), we have that

$$V_\infty(\xi_{n_k-1}^k) - V_\infty(x_i) \leq \mathcal{B}(x_i, \xi_{n_k-1}^k) \leq \mathcal{S}_{N_k}(\xi^k) \quad (\text{A.3.31a})$$

$$V_\infty(x_j) - V_\infty(\xi_{n_k}^k) \leq \mathcal{B}(x_j, \xi_{n_k}^k) \leq \mathcal{S}_{N_k}(\xi^k), \quad (\text{A.3.31b})$$

so, taking the limit $k \rightarrow \infty$ yields

$$V_\infty(x) - V_\infty(x_i) \leq 0 \quad (\text{A.3.32a})$$

$$V_\infty(x_j) - V_\infty(z) \leq 0. \quad (\text{A.3.32b})$$

However, the fact that x_i and x_j are equivalent and [Lemma A.3.7](#) imply that $V_\infty(x_i) = V_\infty(x_j)$ so that $V_\infty(x) \leq V_\infty(z)$. Since α_∞ is increasing, we have that $f(x) \leq f(z)$.

But we have that $\rho(x, z) = 0$ so that $z = \Theta_1(x)$. Therefore, if $\nabla f(x) \neq 0$, we would have

$$f(z) - f(x) = - \int_0^1 \|\nabla f(\Theta_t(x))\|^2 dt < 0, \quad (\text{A.3.33})$$

which would be a contradiction. Therefore, $\nabla f(x) = 0$ and $x = z$. Since x is a critical point, to show that it belongs to $\mathcal{K}$, it suffices to show that $x_i \sim x$.

Take $\delta > 0$. By [Lemma A.3.9](#), $\mathcal{W}_\delta(\{x\})$ is a neighborhood of x . Since $\xi_{n_k}^k$ converges to x , for k large enough, $\xi_{n_k}^k$ belongs to $\mathcal{W}_\delta(\{x\})$. $(\xi_0^k, \dots, \xi_{n_k}^k, x)$ and $(x, \xi_{n_k}^k, \dots, \xi_{N_k-1}^k)$ are, respectively, paths from x_i to x and from x to x_j with action cost of at most $\mathcal{S}_{N_k}(\xi^k) + \delta$ so that, for k large enough, $\mathcal{B}(x_i, x) \leq 2\delta$ and $\mathcal{B}(x, x_j) \leq 2\delta$.

Hence, we have shown that $\mathcal{B}(x_i, x) = \mathcal{B}(x, x_j) = 0$. Then, we also have

$$0 \leq \mathcal{B}(x, x_i) \leq \mathcal{B}(x, x_j) + \mathcal{B}(x_j, x_i) = 0, \quad (\text{A.3.34})$$

and therefore $x \sim x_i$ and thus x belongs to $\mathcal{K}$. This is a contradiction.

Therefore, there must exist some sequence $\xi^{i,j} \in (\mathbb{R}^d)^{N_{i,j}}$ with $\xi_0^{i,j} = x_i$, $\xi_{N_{i,j}-1}^{i,j} = x_j$ such that $\mathcal{S}_{N_{i,j}}(\xi^{i,j}) < \varepsilon$ and $\max_{0 \leq n < N_{i,j}} d(\xi_n^{i,j}, \mathcal{K}) < \varepsilon$.

Finally, for the general class, consider $x, z \in \mathcal{K}$. By (A.3.29), there exists $i, j \in I$ such that $x \in \mathbb{B}(x_i, \delta_{x_i})$, $z \in \mathbb{B}(x_j, \delta_{x_j})$. Consider $\xi \in (\mathbb{R}^d)^{N_{i,j}+2}$ a extension of $\xi^{i,j}$ defined by

$$\xi_n = \begin{cases} x & \text{if } n = 0, \\ \xi_{n-1}^{i,j} & \text{if } 0 < n \leq N_{i,j}, \\ z & \text{if } n = N_{i,j} + 1, \end{cases} \quad (\text{A.3.35})$$

that still satisfies $\max_{0 \leq n < N_{i,j}+2} d(\xi_n, \mathcal{K}) < \varepsilon$. Then, one has that

$$\mathcal{S}_{N_{i,j}+2}(\xi) < \mathcal{S}_{N_{i,j}}(\xi^{i,j}) + 2\varepsilon \leq 3\varepsilon, \quad (\text{A.3.36})$$

which concludes the proof. $\blacksquare$

The following lemma is inspired by Alongi and Nelson (2007, Prop. 3.3.11).

Lemma A.3.11. *Equivalence classes are connected.*

Proof. Fix $\mathcal{K}$ an equivalence class of $\sim$.

For the sake of contradiction, assume that there are $\mathcal{U}, \mathcal{V}$ disjoint open sets of $\mathbb{R}^d$ such that both $\mathcal{U} \cap \mathcal{K}$ and $\mathcal{V} \cap \mathcal{K}$ are non-empty and $\mathcal{K} = (\mathcal{U} \cup \mathcal{V}) \cap \mathcal{K}$.

Take $x \in \mathcal{U} \cap \mathcal{K}$ and $x' \in \mathcal{V} \cap \mathcal{K}$. By Lemma A.3.10, there exists a sequence of paths $\xi^k \in (\mathbb{R}^d)^{N_k}$ with $\xi_0^k = x$, $\xi_{N_k-1}^k = x'$ for $k \geq 0$ such that $\mathcal{S}_{N_k}(\xi^k) \rightarrow 0$ as $k \rightarrow \infty$ and $\max_{0 \leq n < N_k} d(\xi_n^k, \mathcal{K}) \rightarrow 0$ as $k \rightarrow \infty$. Define a_k as the last point of ξ^k that belongs to $\mathcal{U}$ and b_k as the successor of a_k in ξ^k . Formally, a_k and b_k are defined by $a_k = \xi_{i_k}^k$ and $b_k = \xi_{i_k+1}^k$ where $i_k = \max\{i < N_k : \xi_i^k \in \mathcal{U}\}$.

By construction, both $d(a_k, \mathcal{K})$ and $d(b_k, \mathcal{K})$ go to zero as $k \rightarrow \infty$. In particular, both sequences lie in $\mathcal{U}_1(\mathcal{K})$ from some point onward, which is relatively compact, so they admit convergent subsequences. Without loss of generality, thus assume that $a_k \rightarrow a$ and $b_k \rightarrow b$ as $k \rightarrow \infty$ and a, b belong to $\mathcal{K}$.

Since a_k belongs to $\mathcal{U}$ for all $k \geq 0$, a_k is never in $\mathcal{V}$ so that a does not belong to $\mathcal{V}$ either. Since a belongs to $\mathcal{K}$, it must thus belong to $\mathcal{U}$. Similarly, b must belong to $\mathcal{V}$.

However, by construction, we have that $\rho(a_k, b_k) \leq \mathcal{S}_{N_k}(\xi^k) \rightarrow 0$ as $k \rightarrow \infty$ so that $\rho(a_k, b_k)$ converges to 0 as $k \rightarrow \infty$ as well. By l.s.c. of ρ (Corollary A.2.9 with $N = 1$), $\rho(a, b) = 0$ so that $b = F(a)$ by Lemma A.3.2. But since a belongs to $\mathcal{K}$, it is a critical point of f and therefore $a = F(a)$. Hence $a = b$ with $a \in \mathcal{U}$ and $b \in \mathcal{V}$, which is a contradiction. $\blacksquare$

Lemma A.3.12. *Any connected component of $\text{crit } f$ is included in a single equivalence class.*

Proof. Let C be a connected component of $\text{crit } f$ and fix $x, x' \in C$. We begin by considering a stronger version of Assumption A.1(c), namely that there exists $\gamma \in \mathcal{C}([0, 1], C)$ absolutely continuous such that $\gamma_0 = x$, $\gamma_1 = x'$, i.e., such that γ is differentiable almost everywhere with $\int_0^1 \|\dot{\gamma}_t\| dt < \infty$. We show that $x \sim x'$, i.e., that $\mathcal{B}(x, x') = \mathcal{B}(x', x) = 0$.

Let us begin by showing that $\mathcal{B}(x, x') = 0$.

Since $\text{crit } f$ is compact and C is closed as connected component of a closed set, C is compact. By invoking Lemma A.3.4 at every point of C and extracting

a finite covering from the family of balls obtained, we have that, there exists $\delta > 0$, $L > 0$ such that, for every $x \in C$, $v \in \mathbb{B}(0, \delta)$,

$$\mathcal{L}(x, v) \leq L\|v\|^2. \quad (\text{A.3.37})$$

Fix $\varepsilon \in (0, \delta)$ and define, for any $t \in [0, 1]$,

$$\lambda_t := \int_0^t \frac{\max(\|\dot{\gamma}_s\|, 1)}{\varepsilon} ds. \quad (\text{A.3.38})$$

We have that λ is an increasing bijection from $[0, 1]$ to $[0, \lambda_1]$ and is absolutely continuous with $\dot{\lambda}_t = \max(\|\dot{\gamma}_t\|, 1)/\varepsilon$ almost everywhere. Consider $\tau: [0, \lambda_1] \rightarrow [0, 1]$ the inverse of λ , which is absolutely continuous with $\dot{\tau}_t = \varepsilon/\max(\|\dot{\gamma}_{\tau_t}\|, 1)$ almost everywhere. Define $\varphi \in \mathcal{C}([0, \lambda_1], \mathbb{R}^d)$ by $\varphi_t = \gamma_{\tau_t}$ for any $t \in [0, \lambda_1]$. Then, φ is absolutely continuous and, for any $t \in [0, \lambda_1]$,

$$\dot{\varphi}_t = \dot{\gamma}_{\tau_t} \dot{\tau}_t = \frac{\varepsilon \dot{\gamma}_{\tau_t}}{\max(\|\dot{\gamma}_{\tau_t}\|, 1)} \quad (\text{A.3.39})$$

which has norm less than $\varepsilon < \delta$.

Therefore, we have that,

$$\begin{aligned} \mathcal{S}_{0, \lambda_1}(\varphi) &= \int_0^{\lambda_1} \mathcal{L}(\varphi_t, \dot{\varphi}_t) dt \\ &\leq L \int_0^{\lambda_1} \|\dot{\varphi}_t\|^2 dt \\ &\leq \varepsilon L \int_0^{\lambda_1} \|\dot{\varphi}_t\| dt \\ &= \varepsilon L \int_0^{\lambda_1} \dot{\tau}_t \|\dot{\gamma}_{\tau_t}\| dt \\ &= \varepsilon L \int_0^1 \|\dot{\gamma}_t\| dt \end{aligned} \quad (\text{A.3.40})$$

where the last equality is obtained by the change of variable $t \leftarrow \tau_t$. Thus, we have shown that, for any $\varepsilon \in (0, \delta)$,

$$\mathcal{B}(x, x') \leq \varepsilon L \int_0^1 \|\dot{\gamma}_t\| dt \quad (\text{A.3.41})$$

with $\int_0^1 \|\dot{\gamma}_t\| dt < \infty$ by construction so that $\mathcal{B}(x, x') = 0$.

Reversing the roles of x and x' and considering the path $(\gamma_{1-t})_{t \in [0, 1]}$ then yields that $\mathcal{B}(x', x) = 0$. Therefore, we have shown that if there exists $\gamma \in \mathcal{C}([0, 1], C)$ absolutely continuous such that $\gamma_0 = x$, $\gamma_1 = x'$, then $x \sim x'$.

We now relax our assumption on the paths from absolute continuity to piecewise absolute continuity ([Assumption A.1](#)). For $x, x' \in C$, by assumption, there exists $\gamma \in \mathcal{C}([0, 1], C)$ such that $\gamma_0 = x$, $\gamma_1 = x'$ and such that it is piecewise absolutely continuous: γ is differentiable almost everywhere and there exists $0 = t_0 < t_1 < \dots < t_N = 1$ such that $\dot{\gamma}$ is integrable on every closed interval of (t_n, t_{n+1}) for $n = 0, \dots, N-1$. Take $0 \leq n < N-1$ and $t_n < s < t < t_{n+1}$. γ restricted to $[s, t]$ is absolutely continuous so that, by the previous case, all the points of $\{\gamma_u : u \in [s, t]\}$ are included in a single equivalence class $\mathcal{K}$. Taking $s \rightarrow t_n$ and $t \rightarrow t_{n+1}$ yields that $\{\gamma_u : u \in (t_n, t_{n+1})\}$ is included in $\mathcal{K}$. Moreover, by continuity of γ , γ_{t_n} and $\gamma_{t_{n+1}}$ belong to the closure of $\mathcal{K}$, which is closed by

Lemma A.3.8, so that γ_{t_n} and $\gamma_{t_{n+1}}$ belong to $\mathcal{K}$ as well. Therefore, $\gamma_{t_n} \sim \gamma_{t_{n+1}}$. By transitivity, we obtain that $x = \gamma_0 \sim \gamma_{t_1} \sim \cdots \sim \gamma_{t_{N-1}} \sim \gamma_1 = x'$ so that $x \sim x'$. ■

Combining **Lemma A.3.11** and **Lemma A.3.12**, we have shown that any connected component of $\text{crit } f$ is included in a single equivalence class and since they are connected, two distinct connected components of $\text{crit } f$ cannot belong to the same equivalent class; hence, we have that they coincide.

Corollary A.3.13. *The equivalence classes of $\sim$ are exactly connected components of $\text{crit } f$.*

We end this section by providing a sufficient condition for $\mathcal{B}(x, x')$ to be finite.

Lemma A.3.14. *Consider $x, x' \in \mathbb{R}^d$ and assume that there exists $T > 0$, $\gamma \in \mathcal{C}([0, T], \mathbb{R}^d)$ absolutely continuous such that $\gamma_0 = x$, $\gamma_T = x'$ and such that, for every $t \in [0, T]$, $\nabla f(\gamma_t)$ is in the interior of the closed convex hull of the support of $U(\gamma_t, \omega)$, i.e.,*

$$\nabla f(\gamma_t) \in \text{int } \overline{\text{conv}} \text{ supp } U(\gamma_t, \omega). \quad (\text{A.3.42})$$

Then, $\mathcal{B}(x, x') < \infty$.

Proof. First, fix $t \in [0, T]$ and consider the exponential family of distributions

$$\propto e^{\langle p, U(\gamma_t, \omega) \rangle} \mathbb{P}(d\omega) \quad (\text{A.3.43})$$

with parameter $p \in \mathbb{R}^d$. We begin by applying **Brown (1986, Thm. 3.6)**. The family above is indeed minimal in the sense of **Brown (1986, §1.1)** since the interior of $\overline{\text{conv}} \text{ supp } U(\gamma_t, \omega)$ is non-empty by assumption. Combined with (A.3.42), **Brown (1986, Thm. 3.6)** then implies that, for every $t \in [0, T]$, $\nabla f(\gamma_t)$ belongs to $\nabla_p \bar{\mathcal{H}}(\gamma_t, \mathbb{R}^d)$ so there exists $p_t \in \mathbb{R}^d$ such that $\nabla f(\gamma_t) = \nabla_{p_t} \bar{\mathcal{H}}(\gamma_t, p_t)$. As in the proof of **Lemma A.3.4**, we will now invoke the implicit function theorem on the equation

$$\nabla_p \bar{\mathcal{H}}(x, p) = \nabla f(x) + v. \quad (\text{A.3.44})$$

For this, need to check that $\text{Hess}_p \bar{\mathcal{H}}(\gamma_t, p_t)$ is positive definite. For notational convenience, denote by $\mathbb{E}_{p_t}$, cov_{p_t} and $\mathbb{V}_{p_t}$ the expectation, covariance and variance operators with respect to the reweighted distribution

$$\propto e^{\langle p_t, U(\gamma_t, \omega) \rangle} \mathbb{P}(d\omega). \quad (\text{A.3.45})$$

For any non-zero vector $q \in \mathbb{R}^d$, computing the Hessian of $\bar{\mathcal{H}}$ at (γ_t, p_t) yields that

$$\langle q, \text{Hess}_p \bar{\mathcal{H}}(\gamma_t, p_t) q \rangle = \langle q, \text{cov}_{p_t}[U(\gamma_t, \omega)] q \rangle \quad (\text{A.3.46})$$

$$= \mathbb{V}_{p_t}[\langle q, U(\gamma_t, \omega) \rangle] \quad (\text{A.3.47})$$

$$= \min_{\alpha \in \mathbb{R}} \mathbb{E}_{p_t} [(\langle q, U(\gamma_t, \omega) \rangle - \alpha)^2] \quad (\text{A.3.48})$$

$$= \min_{\alpha \in \mathbb{R}} \frac{\mathbb{E}[(\langle q, U(\gamma_t, \omega) \rangle - \alpha)^2 e^{\langle p_t, U(\gamma_t, \omega) \rangle}]}{\mathbb{E}[e^{\langle p_t, U(\gamma_t, \omega) \rangle}]} \quad (\text{A.3.49})$$

The blanket assumptions on the noise ([Appendix A.1.2](#)) and Cauchy-Schwarz inequality then yield

$$\langle q, \text{Hess}_p \tilde{\mathcal{H}}(\gamma_t, p_t) q \rangle \geq e^{-2\|p_t\|^{M(1+\|\gamma_t\|)}} \min_{\alpha \in \mathbb{R}} \mathbb{E} \left[(\langle q, U(\gamma_t, \omega) \rangle - \alpha)^2 \right] \quad (\text{A.3.50})$$

$$= e^{-2\|p_t\|^{M(1+\|\gamma_t\|)}} \langle q, \text{cov}[U(\gamma_t, \omega)] q \rangle, \quad (\text{A.3.51})$$

which is positive by the positive definiteness of $\text{cov}[U(\gamma_t, \omega)]$ ([Appendix A.1.2](#)). Hence, $\text{Hess}_p \tilde{\mathcal{H}}(\gamma_t, p_t)$ is positive definite and we can apply the implicit function theorem to ([A.3.44](#)) to obtain that there exists $\delta(\gamma_t) > 0$, $p: \mathbb{B}(\gamma_t, \delta(\gamma_t)) \times \mathbb{B}(0, \delta(\gamma_t)) \rightarrow \mathbb{R}^d$ such that, for any $x \in \mathbb{B}(\gamma_t, \delta(\gamma_t))$, $v \in \mathbb{B}(0, \delta(\gamma_t))$,

$$\nabla_p \tilde{\mathcal{H}}(x, p(x, v)) = \nabla f(x) + v, \quad (\text{A.3.52})$$

or, equivalently

$$\nabla_p \mathcal{H}(x, p(x, v)) = v. \quad (\text{A.3.53})$$

Therefore, as in the proof of [Lemma A.3.4](#), we obtain that $\mathcal{L}$ is continuous on $\overline{\mathbb{B}}(\gamma_t, \delta(\gamma_t)/2) \times \overline{\mathbb{B}}(0, \delta(\gamma_t)/2)$ and therefore bounded by $M(\gamma_t) > 0$. Since γ is continuous, $\{\gamma_t : t \in [0, T]\}$ is compact and therefore, by extracting a finite covering from

$$\bigcup_{t \in [0, T]} \mathbb{B}(\gamma_t, \delta(\gamma_t)/2), \quad (\text{A.3.54})$$

we obtain that there exists $\delta > 0$ and $M > 0$ such that, for every $t \in [0, T]$, $\mathcal{L}(\gamma_t, \cdot)$ is finite and bounded by M on $\mathbb{B}(0, \delta)$. Choosing a φ reparametrization of γ , which is $\mathcal{C}^1$, such that $\|\dot{\varphi}_t\| < \delta$ for every $t \in [0, S]$, we thus obtain a path such that

$$S_{0,S}(\varphi) = \int_0^S \mathcal{L}(\varphi_t, \dot{\varphi}_t) dt \leq MS < \infty, \quad (\text{A.3.55})$$

which implies that $\mathcal{B}(x, x') < \infty$. ■

A.3.3 Lyapunov condition

Definition A.4 (Stopping times for the accelerated process). For any set $A \subset \mathbb{R}^d$, we define the hitting and exit times of A :

$$\tau_A^\eta := \inf\{n \geq 1 : x_n^\eta \in A\} \quad (\text{A.3.56a})$$

$$\sigma_A^\eta := \inf\{n \geq 0 : x_n^\eta \notin A\}. \quad (\text{A.3.56b})$$

We will need the following concentration lemma and its corollary.

Lemma A.3.15 (Part of the proof of ([Rigollet and Hütter, 2023](#), Th. 1.19)). Let X be a random variable in $\mathbb{R}^d$ such that, for all $v \in \mathbb{R}^d$,

$$\log \mathbb{E}[\exp(\langle v, X \rangle)] \leq \frac{\|v\|^2}{2}. \quad (\text{A.3.57})$$

Then, for all $t > 0$, we have

$$\mathbb{P}(\|X\|^2 \geq t) \leq 6^d \exp\left(-\frac{t}{8}\right). \quad (\text{A.3.58})$$

Corollary A.3.16. *In the context of Lemma A.3.15, it holds that*

$$\mathbb{E}[\|X\|^2] \leq 16d \log 6. \quad (\text{A.3.59})$$

Proof. Since $\|X\|^2$ is non-negative, its expectation can be written as

$$\begin{aligned} \mathbb{E}[\|X\|^2] &= \int_0^{+\infty} \mathbb{P}(\|X\|^2 > t) dt \\ &\leq 8d \log 6 + \int_{8d \log 6}^{+\infty} \mathbb{P}(\|X\|^2 > t) dt. \end{aligned} \quad (\text{A.3.60})$$

Invoking Lemma A.3.15 then yields that

$$\mathbb{E}[\|X\|^2] \leq 8d \log 6 + 6^d \int_{8d \log 6}^{+\infty} \exp\left(-\frac{t}{8}\right) dt \leq 8d \log 6 + 8, \quad (\text{A.3.61})$$

which concludes the proof since $1 \leq d \log 6$. $\blacksquare$

Lemma A.3.17 (Lyapunov condition). *Define V_∞ as in Lemma A.3.7. Then, there exists $\mathcal{K} \subset \mathbb{R}^d$ compact, $\eta_0 > 0$, $c > 0$ such that, for any $\eta \leq \eta_0$, $n \geq 0$, if $x_n \notin \mathcal{K}$, then, almost surely,*

$$\begin{aligned} V_\infty(x_{n+1}) - V_\infty(x_n) &\leq \eta \left(\frac{\|U(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} \right) \\ &\leq \eta \left(\frac{\|U(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - (16d \log 6 + c) \right). \end{aligned} \quad (\text{A.3.62})$$

Proof. By Assumption A.3, there is $R \geq \frac{1}{2}$, $c > 0$ such that, for any $x \in \mathbb{R}^d$ such that $\|x\| \geq R$,

$$\begin{cases} f(x) \geq c \\ c \leq \frac{\sigma_\infty^2(f(x))}{\|x\|^2} \leq c^{-1} \\ \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \geq 16d \log 6 + c. \end{cases} \quad (\text{A.3.63})$$

Then, define $\mathcal{K} := \overline{\mathbb{B}}(0, 2R + 1)$.

By definition, V_∞ is twice continuously differentiable and, its Hessian satisfies, for any $x \in \mathbb{R}^d$,

$$\text{Hess } V_\infty(x) \leq \frac{2 \text{Hess } f(x)}{\sigma_\infty^2(f(x))} \leq \frac{2\beta(f)}{\sigma_\infty^2(f(x))} I_d. \quad (\text{A.3.64})$$

For the sake of clarity, for any $n \geq 0$, denote by δx_n the quantity

$$\delta x_n = \frac{x_{n+1} - x_n}{\eta}. \quad (\text{A.3.65})$$

For any $n \geq 0$, we now have

$$\begin{aligned} V_\infty(x_{n+1}) - V_\infty(x_n) &\leq \eta \langle \nabla V_\infty(x_n), \delta x_n \rangle \\ &\quad + \frac{\eta^2}{2} \|\delta x_n\|^2 \sup_{t \in [0,1]} \frac{2\beta(f)}{\sigma_\infty^2(f(x_n + t(x_{n+1} - x_n)))}. \end{aligned} \quad (\text{A.3.66})$$

We first focus on bounding the last term. First note that, by the blanket assumptions, $\|x_{n+1} - x_n\| \leq 2\eta M(1 + \|x_n\|)$ so that, For any $t \in [0, 1]$, $\eta \leq (4M)^{-1}$,

$$\begin{aligned} \|x_n + t(x_{n+1} - x_n)\| &\geq \|x_n\| - \|x_{n+1} - x_n\| \\ &\geq \|x_n\| - 2\eta M(1 + \|x_n\|) \\ &\geq \frac{1}{2}(\|x_n\| - 1). \end{aligned} \quad (\text{A.3.67})$$

If x_n is outside of $\mathcal{K}$, we have that $\|x_n\| \geq 2R + 1$ and thus $\|x_n + t(x_{n+1} - x_n)\| \geq R$. Moreover, since $R \geq \frac{1}{2}$, x_n being outside of $\mathcal{K}$ also implies that $\frac{1}{2}\|x_n\| \geq 1$ and so $\|x_n + t(x_{n+1} - x_n)\| \geq \frac{1}{4}\|x_n\|$. By the definition of R , we thus have

$$\begin{aligned} \sigma_\infty^2(f(x_n + t(x_{n+1} - x_n))) &\geq c\|x_n + t(x_{n+1} - x_n)\|^s \\ &\geq c\left(\frac{1}{4}\right)^s \|x_n\|^s \\ &\geq \frac{c^2}{4^s} \sigma_\infty^2(f(x_n)). \end{aligned} \quad (\text{A.3.68})$$

Thus, if $x_n \notin \mathcal{K}$, (A.3.64) yields that

$$V_\infty(x_{n+1}) - V_\infty(x_n) \leq \eta \langle \nabla V_\infty(x_n), \delta x_n \rangle + \eta^2 \frac{4^s \beta(f)}{c^2 \sigma_\infty^2(f(x_n))} \|\delta x_n\|^2. \quad (\text{A.3.69})$$

Now, we can rewrite the inner product as

$$\begin{aligned} \langle \nabla V_\infty(x_n), \delta x_n \rangle &= \frac{2\langle \nabla f(x_n), \delta x_n \rangle}{\sigma_\infty^2(f(x_n))} \\ &= \frac{\|\delta x_n + \nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\delta x_n\|^2}{\sigma_\infty^2(f(x_n))}. \end{aligned} \quad (\text{A.3.70})$$

Plugging this into (A.3.69) and assuming that $\eta \leq \frac{c^2}{4^s \beta(f)}$, we obtain

$$\begin{aligned} V_\infty(x_{n+1}) - V_\infty(x_n) &\leq \eta \left(\frac{\|\delta x_n + \nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} \right) \\ &\leq \eta \left(\frac{\|U(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - \frac{\|\nabla f(x_n)\|^2}{\sigma_\infty^2(f(x_n))} \right) \end{aligned} \quad (\text{A.3.71})$$

where we unrolled δx_n in the last inequality. Using again that $x_n \notin \mathcal{K}$, we obtain

$$V_\infty(x_{n+1}) - V_\infty(x_n) \leq \eta \left(\frac{\|U(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - (16d \log 6 + c) \right) \quad (\text{A.3.72})$$

which concludes the proof. $\blacksquare$

We will reuse, in later sections, the following fact that we thus state as a lemma: the sequence of iterates of SGD is (*weak*) Feller (see e.g., (Hernández-Lerma and Lasserre, 2003, Def. 4.4.2)).

Lemma A.3.18. *The Markov chain $(x_n)_{n \geq 0}$ is (*weak*) Feller.*

Proof. since both ∇f and U are continuous, for any $g: \mathbb{R}^d \rightarrow \mathbb{R}$ continuous and bounded, the function.

$$x \in \mathbb{R}^d \mapsto \mathbb{E}_x[g(x_1)] = \mathbb{E}_x[g(x - \eta \nabla f(x) + \eta U(x, \omega))] \quad (\text{A.3.73})$$

is still continuous and bounded. Therefore, the Markov chain $(x_n)_{n \geq 0}$ is weak-Feller. ■

Lemma A.3.19. *There is $\eta_0 > 0$ such that, for any $\eta \leq \eta_0$, there exists an invariant probability measure for $(x_n)_{n \geq 0}$.*

Proof. We invoke a general result on weak-Feller Markov chains that satisfy a Lyapunov condition, e.g., [Hernández-Lerma and Lasserre \(2003, Thm. 7.2.4\)](#) or [Douc et al. \(2018, Thm. 12.3.6\)](#).

First, [Lemma A.3.18](#) ensures that $(x_n)_{n \geq 0}$ is weak-Feller.

Moreover, by [Lemma A.3.17](#), there exists $\mathcal{K} \subset \mathbb{R}^d$ compact, $\eta_0 > 0$, $c > 0$ such that, for any $\eta \leq \eta_0$, $x_0 = x \notin \mathcal{K}$,

$$V_\infty(x_1) - V_\infty(x) \leq \eta \left(\frac{\|U(x, \omega_0)\|^2}{\sigma_\infty^2(f(x))} - (16d \log 6 + c) \right). \quad (\text{A.3.74})$$

Passing to the expectation yields that, for any $x \notin \mathcal{K}$,

$$\mathbb{E}_x[V_\infty(x_1)] - V_\infty(x) \leq \eta \left(\frac{\mathbb{E}_x[\|U(x, \omega_0)\|^2]}{\sigma_\infty^2(f(x))} - (16d \log 6 + c) \right). \quad (\text{A.3.75})$$

Applying [Corollary A.3.16](#) with $X \leftarrow \frac{U(x, \omega_0)}{\sqrt{\sigma_\infty^2(f(x))}}$ (the conditions of application are verified from [Assumption A.2\(c\)](#)) yields that

$$\mathbb{E}_x[V_\infty(x_1)] - V_\infty(x) \leq -\eta c. \quad (\text{A.3.76})$$

Hence, for any $x \in \mathbb{R}^d$, it holds

$$\mathbb{E}_x[V_\infty(x_1)] - V_\infty(x) \leq -\eta c + \mathbb{1}_{\{x \in \mathcal{K}\}} \left(\sup_{x' \in \mathcal{K}} \mathbb{E}_{x'}[V_\infty(x')] - \inf_{\mathbb{R}^d} V_\infty + \eta c \right), \quad (\text{A.3.77})$$

with V_∞ which is not identically equal to its minimum since f is coercive.

Therefore, we can apply [Hernández-Lerma and Lasserre \(2003, Thm. 7.2.4\)](#) with $V_\infty - \inf_{\mathbb{R}^d} V_\infty$, that guarantees that there exists an invariant probability measure for $(x_n)_{n \geq 0}$. ■

Lemma A.3.20. *There exists a compact set $\mathcal{X} \subset \mathbb{R}^d$, $\eta_0 > 0$, such that for any compact set $\mathcal{X}' \subset \mathbb{R}^d$ such that $\mathcal{X} \subset \mathcal{X}'$, there exists $a, b > 0$, such that*

$$\forall \eta \leq \eta_0, x \in \mathcal{X}' \setminus \mathcal{X}, n \geq 0, \mathbb{P}_x(\tau_{\mathcal{X}}^\eta > n) \leq \exp\left(-\frac{an}{\eta} + \frac{b}{\eta}\right). \quad (\text{A.3.78})$$

Proof. This result is a consequence of [Lemma A.3.17](#): there exists $\mathcal{K} \subset \mathbb{R}^d$ compact, $\eta_0 > 0$, $c > 0$ such that, for any $\eta \leq \eta_0$, $n \geq 0$, if $x_n \notin \mathcal{K}$, then, almost surely,

$$V_\infty(x_{n+1}) - V_\infty(x_n) \leq \eta \left(\frac{\|U(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - (16d \log 6 + c) \right). \quad (\text{A.3.79})$$

First, let us choose $\mathcal{X}$. There is some $R > 0$ such that $\mathcal{K} \subset \mathbb{B}(0, R)$. Define $\tilde{R} := e^{4M+1}(R+1)$ and $\mathcal{X} := \mathbb{B}(0, \tilde{R})$. With $\eta \leq (4M)^{-1}$, if, for some $n \geq 0$, x_n is not in $\mathcal{X}$, by [Lemma A.1.3](#), for any $k \leq \lceil \eta^{-1} \rceil$, x_{n+k} has norm greater or equal to R and thus x_{n+k} is not in $\mathcal{K}$ either.

Now, fix $N \geq 0$ and consider the event $\{\tau_{\mathcal{X}}^\eta > N\}$ with $x_0 = x \in \mathcal{X}' \setminus \mathcal{X}$. This means that, for any $0 \leq n \leq N$, x_n^η is outside of $\mathcal{X}$ and so, for any $0 \leq n \leq N \lfloor \eta^{-1} \rfloor$, x_n is not in $\mathcal{K}$.

Summing (A.3.79) over $n = 0, \dots, N \lfloor \eta^{-1} \rfloor - 1$ yields

$$\begin{aligned} V_\infty(x_N^\eta) - V_\infty(x_0) &= V_\infty(x_{N \lfloor \eta^{-1} \rfloor}) - V_\infty(x_0) \\ &\leq \eta \sum_{n=0}^{N \lfloor \eta^{-1} \rfloor - 1} \left(\frac{\|U(x_n, \omega_n)\|^2}{\sigma_\infty^2(f(x_n))} - (16d \log 6 + c) \right). \end{aligned} \quad (\text{A.3.80})$$

Define $\Delta := \sup_{\mathcal{X}' \setminus \mathcal{X}} V_\infty - \inf_{\mathbb{R}^d \setminus \mathcal{X}} V_\infty$ which is finite since f is coercive. By definition, we have that $V_\infty(x_N^\eta) - V_\infty(x_0) \geq -\Delta$.

Therefore, on the event $\{\tau_{\mathcal{X}}^\eta > N\}$, we have

$$\sum_{n=0}^{N \lfloor \eta^{-1} \rfloor - 1} \frac{\|U(x_n, \omega_n)\|^2}{\sigma_\infty^2 \circ f(x_n)} \geq N \lfloor \eta^{-1} \rfloor (16d \log 6 + c) - \frac{\Delta}{\eta}. \quad (\text{A.3.81})$$

Now, on the whole event, consider the random variable $X \in \mathbb{R}^{N \lfloor \eta^{-1} \rfloor d}$ defined by

$$(X_{nd} + 1, \dots, X_{(n+1)d}) = \frac{U(x_n, \omega_n)}{\sqrt{\sigma_\infty^2 \circ f(x_n)}} \quad \text{for } n = 0, \dots, N \lfloor \eta^{-1} \rfloor - 1, \quad (\text{A.3.82})$$

we have that $\mathbb{P}_x(\tau_{\mathcal{X}}^\eta > N) \leq \mathbb{P}_x(\|X\|^2 \geq N \lfloor \eta^{-1} \rfloor (16d \log 6 + c) - \frac{\Delta}{\eta})$.

Since, for any $v \in \mathbb{R}^d$,

$$\log \mathbb{E} \left[\exp \left(\left\langle v, \frac{U(x_n, \omega_n)}{\sqrt{\sigma_\infty^2 \circ f(x_n)}} \right\rangle \right) \middle| x_0, x_1, \dots, x_n \right] \leq \frac{\|v\|^2}{2}, \quad (\text{A.3.83})$$

the random variable X satisfies the assumptions of [Lemma A.3.15](#) with $d \leftarrow N \lfloor \eta^{-1} \rfloor d$.

First, suppose that $t := N \lfloor \eta^{-1} \rfloor (16d \log 6 + c) - \frac{\Delta}{\eta}$ is non-negative. Applying [Lemma A.3.15](#) with this t yields that

$$\begin{aligned} \mathbb{P}_x(\tau_{\mathcal{X}}^\eta > N) &\leq \mathbb{P}_x \left(\|X\|^2 \geq N \lfloor \eta^{-1} \rfloor (16d \log 6 + c) - \frac{\Delta}{\eta} \right) \\ &\leq 6^{N \lfloor \eta^{-1} \rfloor d} \exp \left(- \frac{N \lfloor \eta^{-1} \rfloor (16d \log 6 + c) - \frac{\Delta}{\eta}}{8} \right) \\ &= \exp \left(- \frac{N \lfloor \eta^{-1} \rfloor (8d \log 6 + c) - \frac{\Delta}{\eta}}{8} \right). \end{aligned} \quad (\text{A.3.84})$$

If t , defined above, is negative, then in particular $\frac{\Delta}{\eta} \geq N \lfloor \eta^{-1} \rfloor c$ so that this bounds still (trivially) holds.

Finally, in particular, for $\eta \leq 1/2$, $\lfloor \eta^{-1} \rfloor \geq (2\eta)^{-1}$ so we obtain

$$\mathbb{P}_x(\tau_{\mathcal{X}}^\eta > N) \leq \exp \left(- \frac{N(8d \log 6 + c)}{16\eta} + \frac{\Delta}{8\eta} \right) \quad (\text{A.3.85})$$

and our proof is complete. $\blacksquare$

A.3.4 Preliminary estimates and lemmas

We will use the following lemma, which corresponds to [Kifer \(1988a, Lem. 5.3\)](#).

Lemma A.3.21. *Let $\mathcal{K} \subset \mathbb{R}^d$ be a compact set such that $\mathcal{K} \cap \text{crit } f = \emptyset$. Then there exists $c > 0$, $N \geq 1$, $\eta_0 > 0$, such that, for any $n > N$, $x \in \mathcal{K}$, $\eta \leq \eta_0$,*

$$\mathbb{P}_x\left(\tau_{\mathbb{R}^d \setminus \mathcal{K}}^\eta > n\right) = \mathbb{P}_x\left(\sigma_{\mathcal{K}}^\eta > n\right) \leq \exp(-c(n - N)/\eta). \quad (\text{A.3.86})$$

Proof. The proof is exactly the same as the proof [Kifer \(1988a, Lem. 5.3\)](#), which only uses the l.s.c. of $\mathcal{S}_N$ ([Corollary A.2.9](#)). $\blacksquare$

The following lemma provides a convenient reformulation of the results of [Lemma A.3.20](#) and [Lemma A.3.21](#).

Lemma A.3.22.

- *There exists $\mathcal{X} \subset \mathbb{R}^d$ a compact set, $\eta_0 > 0$, such that for any $\mathcal{X}' \subset \mathbb{R}^d$ compact set such that $\mathcal{X} \subset \mathcal{X}'$, there exists $\alpha_0, a, b > 0$ such that,*

$$\forall \eta \leq \eta_0, \alpha \leq \alpha_0, x \in \mathcal{X}' \setminus \mathcal{X}, \quad \mathbb{E}_x \left[e^{\frac{\alpha \tau_{\mathcal{X}}^\eta}{\eta}} \right] \leq e^{\frac{a\alpha}{\eta} + b}. \quad (\text{A.3.87})$$

- *For any $\mathcal{K} \subset \mathbb{R}^d$ compact such that $\mathcal{K} \cap \text{crit } f = \emptyset$, there exists $\eta_0, \alpha_0, a, b > 0$ such that,*

$$\forall \eta \leq \eta_0, \alpha \leq \alpha_0, x \in \mathcal{K}, \quad \mathbb{E}_x \left[e^{\frac{\alpha \sigma_{\mathcal{K}}^\eta}{\eta}} \right] \leq e^{\frac{a\alpha}{\eta} + b}. \quad (\text{A.3.88})$$

Proof. The proofs of both statements are very similar, so we prove only the second one for notational convenience. Fix $\mathcal{K} \subset \mathbb{R}^d$ compact such that $\mathcal{K} \cap \text{crit } f = \emptyset$. By [Lemma A.3.21](#), there exists $c > 0$, $N \geq 1$ such that, for any $n > N$, $x \in \mathcal{K}$, $\eta \leq \eta_0$,

$$\mathbb{P}_x\left(\sigma_{\mathcal{K}}^\eta > n\right) \leq \exp(-c(n - N)/\eta). \quad (\text{A.3.89})$$

Let us first bound $\mathbb{E}_x \left[\exp\left(\frac{\alpha(\sigma_{\mathcal{K}}^\eta - N)}{\eta}\right) \right]$. We have that

$$\begin{aligned} \mathbb{E}_x \left[\exp\left(\frac{\alpha(\sigma_{\mathcal{K}}^\eta - N)}{\eta}\right) \right] &= \int_0^\infty \mathbb{P}_x \left(\exp\left(\frac{\alpha(\sigma_{\mathcal{K}}^\eta - N)}{\eta}\right) > t \right) dt \\ &\leq e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^\infty \mathbb{P}_x \left(\sigma_{\mathcal{K}}^\eta > N + \frac{\eta \log t}{\alpha} \right) dt \\ &\leq e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^\infty \mathbb{P}_x \left(\sigma_{\mathcal{K}}^\eta > N + \left\lfloor \frac{\eta \log t}{\alpha} \right\rfloor \right) dt \\ &\leq e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^\infty \exp\left(-\frac{c}{\eta} \left\lfloor \frac{\eta \log t}{\alpha} \right\rfloor\right) dt, \end{aligned} \quad (\text{A.3.90})$$

where we used [\(A.3.89\)](#) in the last inequality.

Lower bounding $\lfloor s \rfloor$ by $s - 1$, we obtain that

$$\begin{aligned} \mathbb{E}_x \left[\exp \left(\frac{\alpha(\sigma_{\mathcal{K}}^\eta - N)}{\eta} \right) \right] &\leq e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^{\infty} \exp \left(-\frac{c}{\eta} \frac{\eta \log t}{\alpha} + \frac{c}{\eta} \right) dt \\ &= e^{\frac{\alpha}{\eta}} + \int_{e^{\frac{\alpha}{\eta}}}^{\infty} e^{\frac{c}{\eta}} t^{-\frac{c}{\alpha}} dt. \end{aligned} \quad (\text{A.3.91})$$

Performing the change of variable $s \leftarrow e^{-\frac{\alpha}{\eta}} t$, we obtain that

$$\mathbb{E}_x \left[\exp \left(\frac{\alpha(\sigma_{\mathcal{K}}^\eta - N)}{\eta} \right) \right] \leq e^{\frac{\alpha}{\eta}} \left(1 + \int_1^{\infty} s^{-\frac{c}{\alpha}} ds \right). \quad (\text{A.3.92})$$

When $\alpha \leq c/2$, we obtain that

$$\mathbb{E}_x \left[\exp \left(\frac{\alpha(\sigma_{\mathcal{K}}^\eta - N)}{\eta} \right) \right] \leq e^{\frac{\alpha}{\eta}} \left(1 + \int_1^{\infty} s^{-2} ds \right), \quad (\text{A.3.93})$$

and therefore,

$$\mathbb{E}_x \left[\exp \left(\frac{\alpha \sigma_{\mathcal{K}}^\eta}{\eta} \right) \right] \leq e^{\frac{(N+1)\alpha}{\eta}} \left(1 + \int_1^{\infty} s^{-2} ds \right), \quad (\text{A.3.94})$$

which concludes the proof. $\blacksquare$

The following lemma upper-bounds the probability of exiting a large neighborhood of the critical points before visiting a smaller one critical points.

Lemma A.3.23. *Consider $\text{crit } f \subset \mathcal{U} \subset \mathcal{X} \subset \mathbb{R}^d$ with $\mathcal{U}$ an open set and $\mathcal{X}$ a compact set. There exists $\mathcal{X}' \subset \mathbb{R}^d$ compact set such that $\mathcal{X} \subset \mathcal{X}'$, $\Delta > 0$, $\eta_0 > 0$ such that, for any $\eta \leq \eta_0$, $x \in \mathcal{X}$,*

$$\mathbb{P}_x \left(\sigma_{\mathcal{X}'}^\eta < \tau_{\mathcal{U}}^\eta \right) \leq \exp \left(-\frac{\Delta}{\eta} \right). \quad (\text{A.3.95})$$

Proof. Define $\mathcal{X}' := \{x \in \mathbb{R}^d : f(x) \leq \sup_{\mathcal{X}} f + 1\}$ and let V_∞ be as in [Lemma A.3.7](#).

Since α_∞ is (stricly) increasing as its derivative is (stricly) positive by definition, we have

$$\Delta := \alpha_\infty \left(\sup_{\mathcal{X}'} f + 1 \right) - \alpha_\infty \left(\sup_{\mathcal{X}} f \right) > 0. \quad (\text{A.3.96})$$

By [Lemma A.3.21](#) applied to $\mathcal{K} \leftarrow \mathcal{X}'_\delta \setminus \mathcal{U}$, there exists $c > 0$, $N_0 \geq 1$, $\eta_0 > 0$ such that for any $n > N_0$, $x \in \mathcal{X}'_\delta \setminus \mathcal{U}$, $\eta \leq \eta_0$,

$$\mathbb{P}_x \left(\sigma_{\mathcal{X}'_\delta \setminus \mathcal{U}}^\eta > n \right) \leq \exp(-c(n - N_0)/\eta). \quad (\text{A.3.97})$$

Defining $N := \lceil \frac{\Delta}{c} \rceil + N_0$, which is greater or equal than 1, we obtain that, for any $\eta \leq \eta_0$, $x \in \mathcal{X}'_\delta \setminus \mathcal{U}$,

$$\mathbb{P}_x \left(\sigma_{\mathcal{X}'_\delta \setminus \mathcal{U}}^\eta \geq N \right) \leq \exp \left(-\frac{\Delta}{\eta} \right). \quad (\text{A.3.98})$$

Note that this inequality actually holds for any $x \in \mathcal{X}'_\delta$.

We now bound $\mathbb{P}_x(\sigma_{\mathcal{X}'_\delta}^\eta > \tau_{\mathcal{U}}^\eta)$ for $x \in \mathcal{X}$ by distinguishing the cases where $\sigma_{\mathcal{X}'_\delta}^\eta < N$ and $\sigma_{\mathcal{X}'_\delta}^\eta \geq N$. For any $x \in \mathcal{X}$, we have that

$$\begin{aligned} \mathbb{P}_x(\sigma_{\mathcal{X}'_\delta}^\eta < \tau_{\mathcal{U}}^\eta) &\leq \mathbb{P}_x(\sigma_{\mathcal{X}'_\delta}^\eta < \tau_{\mathcal{U}}^\eta, \sigma_{\mathcal{X}'_\delta}^\eta < N) + \mathbb{P}_x(\sigma_{\mathcal{X}'_\delta}^\eta < \tau_{\mathcal{U}}^\eta, \sigma_{\mathcal{X}'_\delta}^\eta \geq N) \\ &\leq \mathbb{P}_x(\sigma_{\mathcal{X}'_\delta}^\eta < N) + \mathbb{P}_x(\sigma_{\mathcal{X}'_\delta \setminus \mathcal{U}}^\eta \geq N) \\ &\leq \mathbb{P}_x(\sigma_{\mathcal{X}'_\delta}^\eta < N) + \exp\left(-\frac{\Delta}{\eta}\right) \end{aligned} \quad (\text{A.3.99})$$

where we used (A.3.98).

We now focus on bounding the first term. For this, we first show that $\sigma_{\mathcal{X}'_\delta}^\eta < N$ implies that

$$\text{dist}_N(x^\eta, \Gamma_N^{\{x\}}(\Delta/2)) > \delta/2. \quad (\text{A.3.100})$$

For the sake of contradiction, suppose that this inequality does not hold. Therefore, there must exist $\xi \in \Gamma_N^{\{x\}}(\Delta/4)$ such that $\text{dist}_N(x^\eta, \xi) < \delta$. In particular, there is some $n < N$ such that $\xi_n \notin \mathcal{X}'$, so that, by Definition A.2,

$$\mathcal{S}_N(\xi) \geq \mathcal{S}_{n+1}(\xi) \geq \mathcal{B}(\xi_0, \xi_n). \quad (\text{A.3.101})$$

By Lemma A.3.7, we have that

$$\begin{aligned} \mathcal{B}(\xi_0, \xi_n) &\geq V_\infty(\xi_n) - V_\infty(\xi_0) \\ &\geq \alpha_\infty\left(\inf_{\mathbb{R}^d \setminus \mathcal{X}'} f\right) - \alpha_\infty\left(\sup_{\mathcal{X}'} f\right), \end{aligned} \quad (\text{A.3.102})$$

since α_∞ is increasing. By construction of $\mathcal{X}'$, we have further that

$$\mathcal{B}(\xi_0, \xi_n) \geq \alpha_\infty\left(\sup_{\mathcal{X}} f + 1\right) - \alpha_\infty\left(\sup_{\mathcal{X}} f\right) = \Delta, \quad (\text{A.3.103})$$

so that $\mathcal{S}_N(\xi) \geq \Delta$, which is a contradiction with $\xi \in \Gamma_N^{\{x\}}(\Delta/2)$.

Therefore, we have that

$$\begin{aligned} \mathbb{P}(\sigma_{\mathcal{X}'_\delta}^\eta < N) &\leq \mathbb{P}(\text{dist}_N(x^\eta, \Gamma_N^{\{x\}}(\Delta/2)) > \delta/2) \\ &\leq \exp\left(-\frac{\Delta}{4\eta}\right) \end{aligned} \quad (\text{A.3.104})$$

where we invoked Corollary A.2.9 with $\delta \leftarrow \delta/2$, $s \leftarrow \Delta/2$, $\varepsilon \leftarrow \Delta/4$. $\blacksquare$

The following lemma is key to our analysis: it shows that SGD spends most of its time near its critical points.

Lemma A.3.24. *Consider $\text{crit } f \subset \mathcal{U} \subset \mathcal{X} \subset \mathbb{R}^d$ with $\mathcal{U}$ an open set and $\mathcal{X}$ a compact set. Then, there is some $\eta_0, \alpha_0, a, b > 0$ such that,*

$$\forall \eta \leq \eta_0, \alpha \leq \alpha_0, x \in \mathcal{X}, \quad \mathbb{E}_x \left[e^{\frac{\alpha \tau_{\mathcal{U}}^\eta}{\eta}} \right] \leq e^{\frac{a\alpha}{\eta} + b}. \quad (\text{A.3.105})$$

Proof. Fix $\varepsilon > 0$. Without loss of generality, assume that $\mathcal{X}$ is large enough to include the compact set given by the first item of Lemma A.3.22 (note that the guarantee of the first item of Lemma A.3.22 still holds even if $\mathcal{X}$ is larger).

Apply [Lemma A.3.23](#) with $\mathcal{U} \leftarrow \mathcal{U}$, $\mathcal{X} \leftarrow \mathcal{X}$ and denote by $\tilde{\mathcal{X}}$ the obtained compact and $\eta_0, \Delta > 0$ such that, for every $\eta \leq \eta_0$, $x \in \mathcal{X}$,

$$\mathbb{P}_x\left(\sigma_{\tilde{\mathcal{X}}}^\eta < \tau_{\mathcal{U}}^\eta\right) \leq \exp\left(-\frac{\Delta}{\eta}\right). \quad (\text{A.3.106})$$

Define $r := \sup_{x \in \tilde{\mathcal{X}}} \|x\|$ and $R = e^{8M}(1+r)$. Assuming that $\eta \leq 1$, by [Lemma A.1.2](#), for any $x \in \tilde{\mathcal{X}}$, the next two iterates of $(x_n^\eta)_{n \geq 0}$ satisfies $\|x_1^\eta\| \leq R$ and $\|x_2^\eta\| \leq R$. Define $\mathcal{X}' := \overline{\mathbb{B}}(0, R)$.

We invoke both items of [Lemma A.3.22](#) with $\mathcal{K} \leftarrow \tilde{\mathcal{X}} \setminus \mathcal{U}$ and denote by $c > 0$ a constant that satisfies the bounds of both items. In the rest of the proof, we consider η and α smaller than the bounds given by this lemma.

Our goal is to bound, for any $N \geq 0$, the quantity,

$$s_N(\alpha, \eta) := \sup_{x \in \mathcal{X}} \mathbb{E}_x \left[\exp\left(\frac{\alpha \tau_{\mathcal{U}}^{\eta N}}{\eta}\right) \right], \quad \text{where} \quad \tau_{\mathcal{U}}^{\eta N} := \min(N, \tau_{\mathcal{U}}^\eta). \quad (\text{A.3.107})$$

Note that, by construction, $s_N(\alpha, \eta)$ is finite.

Take $x \in \mathcal{X}$. In particular, [Lemma A.3.22](#) implies that $\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta < +\infty$ almost surely for all $x \in \mathcal{X}$.

$$\begin{aligned} \exp\left(\frac{\alpha \tau_{\mathcal{U}}^{\eta N}}{\eta}\right) &= \mathbb{1}_{\{x_{\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}^\eta \in \mathcal{U}\}} \exp\left(\frac{\alpha \sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}{\eta}\right) \\ &\quad + \mathbb{1}_{\{x_{\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}^\eta \notin \mathcal{X}_\delta\}} \exp\left(\frac{\alpha \sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}{\eta}\right) \exp\left(\frac{\alpha(\tau_{\mathcal{U}}^{\eta N} - \sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta)}{\eta}\right) \\ &\leq \mathbb{1}_{\{x_{\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}^\eta \in \mathcal{U}\}} \exp\left(\frac{\alpha \sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}{\eta}\right) \\ &\quad + \mathbb{1}_{\{x_{\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}^\eta \notin \mathcal{X}_\delta\}} \exp\left(\frac{\alpha \sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}{\eta}\right) \exp\left(\frac{\alpha \min(\tau_{\mathcal{U}}^\eta - \sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta, N)}{\eta}\right) \end{aligned} \quad (\text{A.3.108})$$

so that we can apply the strong Markov property to the Markov chain $(x_n^\eta)_{n \geq 0}$ with stopping time $\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta$ to obtain that

$$\begin{aligned} \mathbb{E}_x \left[\exp\left(\frac{\alpha \tau_{\mathcal{U}}^{\eta N}}{\eta}\right) \right] &\leq \mathbb{E}_x \left[\mathbb{1}_{\{x_{\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}^\eta \in \mathcal{U}\}} \exp\left(\frac{\alpha \sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}{\eta}\right) \right. \\ &\quad \left. + \mathbb{1}_{\{x_{\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}^\eta \notin \mathcal{X}_\delta\}} \exp\left(\frac{\alpha \sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}{\eta}\right) \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}} \setminus \mathcal{U}}^\eta}^\eta} \left[\exp\left(\frac{\alpha \tau_{\mathcal{U}}^{\eta N}}{\eta}\right) \right] \right] \end{aligned} \quad (\text{A.3.109})$$

We now bound

$$\mathbb{1}\{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta \notin \mathcal{X}_\delta\} \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} \left[\exp\left(\frac{\alpha\tau_{\mathcal{U}}^\eta}{\eta}\right) \right] = \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} \left[\mathbb{1}\{x_0^\eta \notin \mathcal{X}_\delta\} \exp\left(\frac{\alpha\tau_{\mathcal{U}}^\eta}{\eta}\right) \right] \quad (\text{A.3.110})$$

Since x is in $\mathcal{X}$, by definition, $\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta$ is at least equal to 1 and $x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta-1}^\eta$ is still in $\tilde{\mathcal{X}}$. By definition of $\mathcal{X}'$, $x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta$ must still be in $\mathcal{X}'$. Therefore, the guarantee of [Lemma A.3.22](#) applies and, since in particular it implies that $\tau_{\mathcal{X}}^\eta$ is finite almost surely when the chain is started at $x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta$, we can apply the strong Markov property to obtain that

$$\begin{aligned} \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} \left[\mathbb{1}\{x_0 \notin \tilde{\mathcal{X}}\} \exp\left(\frac{\alpha\tau_{\mathcal{U}}^\eta}{\eta}\right) \right] &\leq \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} \left[\mathbb{1}\{x_0 \notin \tilde{\mathcal{X}}\} \exp\left(\frac{\alpha\tau_{\mathcal{X}}^\eta}{\eta}\right) \mathbb{E}_{x_{\tau_{\mathcal{X}}^\eta}^\eta} \left[\exp\left(\frac{\alpha\tau_{\mathcal{U}}^\eta}{\eta}\right) \right] \right] \\ &\leq s_N(\alpha, \eta) \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} \left[\mathbb{1}\{x_0 \notin \tilde{\mathcal{X}}\} \exp\left(\frac{\alpha\tau_{\mathcal{X}}^\eta}{\eta}\right) \right] \end{aligned} \quad (\text{A.3.111})$$

where, for the second inequality, we used the definition of $s_N(\alpha, \eta)$ and the fact that $x_{\tau_{\mathcal{X}}^\eta}$ is in $\mathcal{X}$.

Now, to bound the remaining expectation, note that x_0 does not belong to $\tilde{\mathcal{X}}$ and, *a fortiori*, does not belong to $\mathcal{X}$. Therefore, $\tau_{\mathcal{X}}^\eta$ depends only on $(x_n^\eta)_{n \geq 1}$ and the (weak) Markov property implies that

$$\begin{aligned} \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} \left[\mathbb{1}\{x_0 \notin \tilde{\mathcal{X}}\} \exp\left(\frac{\alpha\tau_{\mathcal{X}}^\eta}{\eta}\right) \right] &\leq \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} \left[\mathbb{1}\{x_0 \notin \tilde{\mathcal{X}}\} \mathbb{E}_{x_1} \left[\exp\left(\frac{\alpha(1+\tau_{\mathcal{X}}^\eta)}{\eta}\right) \right] \right] \\ &\leq \exp\left(\frac{\alpha(1+c)}{\eta}\right) \mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} [\mathbb{1}\{x_0 \notin \tilde{\mathcal{X}}\}], \end{aligned} \quad (\text{A.3.112})$$

by [Lemma A.3.22](#), since the $x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta-1}^\eta$ is still in $\tilde{\mathcal{X}}$ so that x_1 of the chain started at $x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta$ is still in $\mathcal{X}'$.

[Lemma A.3.23](#) then implies that,

$$\mathbb{E}_{x_{\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}^\eta} \left[\mathbb{1}\{x_0 \notin \tilde{\mathcal{X}}\} \exp\left(\frac{\alpha\tau_{\mathcal{X}}^\eta}{\eta}\right) \right] \leq \exp\left(\frac{\alpha(1+a)-\Delta}{\eta} + b\right) \quad (\text{A.3.113})$$

Combining these bounds with [\(A.3.109\)](#) then gives

$$\begin{aligned} \mathbb{E}_x \left[\exp\left(\frac{\alpha\tau_{\mathcal{U}}^\eta}{\eta}\right) \right] &\leq \left(1 + s_N(\alpha, \eta) \exp\left(\frac{\alpha(1+a)-\Delta}{\eta} + b\right) \right) \mathbb{E}_x \left[\exp\left(\frac{\alpha\sigma_{\tilde{\mathcal{X}}\setminus\mathcal{U}}^\eta}{\eta}\right) \right] \\ &\leq \left(1 + s_N(\alpha, \eta) \exp\left(\frac{\alpha(1+a)-\Delta}{\eta} + b\right) \right) \exp\left(\frac{a\alpha}{\eta} + b\right), \end{aligned} \quad (\text{A.3.114})$$

by [Lemma A.3.22](#). Since this inequality is valid for any $x \in \mathcal{X}$, we have shown that

$$s_N(\alpha, \eta) \leq \left(e^{\frac{a\alpha}{\eta} + b} + s_N(\alpha, \eta) e^{\frac{\alpha(1+2a)-\Delta}{\eta} + b} \right). \quad (\text{A.3.115})$$

For $\alpha \leq \Delta/(2(1+2a))$ and η small enough,

$$e^{\frac{\alpha(1+2a)-\Delta}{\eta} + b} \leq e^{-\frac{\Delta}{2\eta} + b} \leq \frac{1}{2}, \quad (\text{A.3.116})$$

and we obtain that

$$s_N(\alpha, \eta) \leq 2e^{\frac{a\alpha}{\eta} + b}. \quad (\text{A.3.117})$$

Taking $N \rightarrow +\infty$ and using Fatou's lemma yields that

$$\sup_{x \in \mathcal{X}} \mathbb{E}_x \left[\exp \left(\frac{\alpha \tau_{\mathcal{U}}^{\eta}}{\eta} \right) \right] \leq 2e^{\frac{a\alpha}{\eta} + b}, \quad (\text{A.3.118})$$

which concludes the proof. $\blacksquare$

A.3.5 Estimates of the invariant measure

Define, for sets $A, B \subset \mathbb{R}^d$,

$$\mathcal{B}(A, B) := \inf \{ \mathcal{B}(x, z) : x \in A, z \in B \}. \quad (\text{A.3.119})$$

Recall that for any i, j , we define

$$\mathcal{B}_{i,j} = \mathcal{B}(\mathcal{K}_i, \mathcal{K}_j) = \inf \{ \mathcal{B}(x, z) : x \in \mathcal{K}_i, z \in \mathcal{K}_j \}. \quad (\text{A.3.120})$$

Recall that, to avoid degenerate cases, we assume that [\(A.3.121\)](#) holds, that is,

$$\mathcal{B}_{ij} < \infty \quad \text{for all } i, j = 1, \dots, K. \quad (\text{A.3.121})$$

With [Lemma A.3.14](#), this assumption is satisfied in particular if the following condition holds: for any $i, j = 1, \dots, K$, there exists a $\mathcal{C}^1$ path γ joining $\mathcal{K}_i$ and $\mathcal{K}_j$ such that, for all t , $\nabla f(\gamma_t)$ belongs to the interior of the closed convex hull of the support of the noise $U(\gamma_t, \omega)$:

$$\nabla f(\gamma_t) \in \text{int } \overline{\text{conv}} \text{ supp } U(\gamma_t, \omega), \quad (\text{A.3.122})$$

This means there is a sufficient level of noise for the probability of going from $\mathcal{K}_i$ to $\mathcal{K}_j$ with at least one path to be non-zero (though it can be vanishingly small). Note that it does not constrain the nature of the noise itself — which can be discrete, continuous or else —, only the support of its distribution.

Moreover, if [\(A.3.121\)](#) did not hold, the same analysis as in this section could still be carried out. We would consider the components of the graph g connected by edges with finite weights proceed with the proof on each of them, and obtain the same results on each of these components. To keep the complexity of the proof reasonable, we will not consider this case here.

We adapt to our context [Kifer \(1988a, Lem. 5.4\)](#) and simplify it using ideas from [Freidlin and Wentzell \(1998b, Chap. 6\)](#).

Definition A.5 (Freidlin and Wentzell (1998b, Chap. 6, §2)). For i, j ,

$$\tilde{\mathcal{B}}_{i,j} := \inf \left\{ S_N(\xi) : N \geq 1, \xi \in (\mathbb{R}^d)^N, \xi_0 \in \mathcal{K}_i, \xi_{N-1} \in \mathcal{K}_j, \xi_n \notin \bigcup_{l \neq i,j} \mathcal{K}_l \text{ for } n = 1, \dots, N-2 \right\}. \quad (\text{A.3.123})$$

We now defined an important object: the law of the (accelerated) iterated at the first time they reach some set $\mathcal{C}^\delta$ (typically a neighborhood of the critical set), following e.g., Douc et al. (2018, Chap. 3.4).

Definition A.6 (Kifer (1988a, Prop. 5.3)). For $\mathcal{C}^\delta$ open set, with hitting time (of the accelerated sequence) $\tau_{\mathcal{C}^\delta}^\eta := \inf\{n \geq 1 : x_n^\eta \in \mathcal{C}^\delta\}$, (as in Definition A.4), denote the law of $x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta$ started at x by $\mathbb{Q}_{\mathcal{C}^\delta}(x, \cdot)$ and the corresponding N -step transition probability by $\mathbb{Q}_{\mathcal{C}^\delta}^N(x, \cdot)$.

In words, $\mathbb{Q}_{\mathcal{C}^\delta}(x, \cdot)$ is the distributions of the x^η started at x at the first time they reach $\mathcal{C}^\delta$. Typically, $\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{U})$ is the probability that $\mathcal{U} \subset \mathcal{C}^\delta$ is reached first among $\mathcal{C}^\delta$. Then, $\mathbb{Q}_{\mathcal{C}^\delta}^N(x, \cdot)$ is the distribution of the x^η started at x at the N -th time they reach $\mathcal{C}^\delta$.

We first give estimates of the transition probabilities using the $\tilde{\mathcal{B}}_{i,j}$. We will then translate them to $\mathcal{B}_{i,j}$.

Lemma A.3.25. For any $\varepsilon > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{K}_i$, $i = 1, \dots, K$, there is some $\eta_0 > 0$ such that for all i, j , $x \in \mathcal{C}_i^\delta$, $0 < \eta < \eta_0$,

$$\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) \leq \exp \left(-\frac{\tilde{\mathcal{B}}_{i,j}}{\eta} + \frac{\varepsilon}{\eta} \right). \quad (\text{A.3.124})$$

where we defined $\mathcal{C}^\delta := \bigcup_{i=1}^K \mathcal{C}_i^\delta$.

Proof. Assume that, without loss of generality, ε is small enough so that Lemma A.3.9 with $r \leftarrow \varepsilon$ can be applied to every $\mathcal{K}_i$, $i = 1, \dots, K$. Denote by $\mathcal{W}_i$, $i = 1, \dots, K$ the corresponding neighborhoods of $\mathcal{K}_i$.

Since these $\mathcal{W}_i$'s are neighborhoods of the $\mathcal{K}_i$'s, there exists $\delta > 0$ such that $\mathcal{U}_{2\delta}(\mathcal{K}_i) \subset \mathcal{W}_i$ for all $i = 1, \dots, K$. Require then that $\mathcal{C}_i^\delta$ be contained in $\mathcal{U}_\delta(\mathcal{K}_i)$ so that $\mathcal{U}_\delta(\mathcal{C}_i^\delta) \subset \mathcal{W}_i$ for all $i = 1, \dots, K$. Moreover, assume that $\delta > 0$ is small enough so that the neighborhoods $\mathcal{U}_\delta(\mathcal{K}_i)$, $i = 1, \dots, K$ are pairwise disjoint.

Define $0 < \delta' \leq \delta$ such that $\mathcal{U}'_\delta(\mathcal{K}_i)$ is contained in $\mathcal{C}_i^\delta$ for all $i = 1, \dots, K$.

Fix $i, j \in I$ and consider $\xi \in (\mathbb{R}^d)^N$ such that $\xi_0 \in \mathcal{U}'_\delta(\mathcal{C}_i^\delta) \subset \mathcal{W}_i$, $\xi_{N-1} \in \mathcal{U}'_\delta(\mathcal{C}_j^\delta) \subset \mathcal{W}_j$ and $\xi_n \in \mathcal{U}_{\delta'}(\mathbb{R}^d \setminus \bigcup_{l \neq i,j} \mathcal{C}_l^\delta)$ for all $n = 1, \dots, N-2$. By the choice of δ' , ξ_n cannot be in $\bigcup_{l \neq i,j} \mathcal{K}_l$ for any $n = 1, \dots, N-2$.

By definition of $\mathcal{W}_i$ and $\mathcal{W}_j$, there are $x \in \mathcal{K}_i$, $x' \in \mathcal{K}_j$ such that $\rho(x, \xi_0) < \varepsilon$, $\rho(\xi_{N-1}, x') < \varepsilon$. Therefore, the path $\zeta \in (\mathbb{R}^d)^{N+2}$ defined as $\zeta = (x, \xi_0, \xi_1, \dots, \xi_{N-1}, x')$ satisfies

$$S_N(\xi) \geq S_{N+2}(\zeta) - 2\varepsilon. \quad (\text{A.3.125})$$

and, by definition of $\tilde{\mathcal{B}}_{i,j}$, we thus obtain,

$$S_N(\xi) \geq S_{N+2}(\zeta) - 2\varepsilon \geq \tilde{\mathcal{B}}_{i,j} - 2\varepsilon. \quad (\text{A.3.126})$$

Fix $x \in \mathcal{C}_i^\delta$. Let us now bound the probability

$$\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) = \mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta \right). \quad (\text{A.3.127})$$

We have, for any $N \geq 0$,

$$\mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_i^\delta \right) \leq \mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_i^\delta, \tau_{\mathcal{C}^\delta}^\eta < N \right) + \mathbb{P}_x \left(\tau_{\mathcal{C}^\delta}^\eta \geq N \right). \quad (\text{A.3.128})$$

We first bound the second probability using [Lemma A.3.24](#) applied to $\mathcal{U} \leftarrow \mathcal{C}_i^\delta$. Take N such that $\alpha_0(a - N) + \eta_0 b \leq -\tilde{\mathcal{B}}_{i,j}$. Then, by Markov's inequality and [Lemma A.3.24](#), it holds that for all $\eta \leq \eta_0$

$$\begin{aligned} \mathbb{P}_x \left(\tau_{\mathcal{C}^\delta}^\eta \geq N \right) &\leq \mathbb{P}_x \left(\exp \left(\frac{\alpha_0 \tau_{\mathcal{C}^\delta}^\eta}{\eta} \right) \geq \exp \left(\frac{\alpha_0 N}{\eta} \right) \right) \\ &\leq \exp \left(\frac{\alpha_0(a - N)}{\eta} + b \right) \\ &\leq \exp \left(\frac{-\tilde{\mathcal{B}}_{i,j}}{\eta} \right). \end{aligned} \quad (\text{A.3.129})$$

We now bound the term $\mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta, \tau_{\mathcal{C}^\delta}^\eta < N \right)$ for this choice of N .

For this, we show that $x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta$ with $\tau_{\mathcal{C}^\delta}^\eta < N$ implies that

$$\text{dist}_N \left(x^\eta, \Gamma_N^{\{x\}} \left(\tilde{\mathcal{B}}_{i,j} - 3\varepsilon \right) \right) > \frac{\delta'}{2}. \quad (\text{A.3.130})$$

Indeed, on the event $x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta$ with $\tau_{\mathcal{C}^\delta}^\eta < N$, there is some $N' \leq N$ such that $\tau_{\mathcal{C}^\delta}^\eta = N' - 1$. If $\text{dist}_N \left(x^\eta, \Gamma_N^{\{x\}} \left(\tilde{\mathcal{B}}_{i,j} - 3\varepsilon \right) \right) > \frac{\delta'}{2}$ did not hold, this would mean that there exists $\xi \in (\mathbb{R}^d)^{N'}$ such that $\text{dist}_{N'}(x^\eta, \xi) < \delta'$, $\xi_0 = x$ and $\mathcal{S}_{N'}(\xi) \leq \tilde{\mathcal{B}}_{i,j} - 3\varepsilon$. In particular, ξ would also satisfy $\xi_{N'-1} \in \mathcal{U}_{\delta'}(\mathcal{C}_j^\delta)$, $\xi_n \in \mathcal{U}_{\delta'}(\mathbb{R}^d \setminus \mathcal{C}^\delta)$ for all $n = 1, \dots, N' - 2$. This would be in direct contradiction of [\(A.3.126\)](#).

Therefore, we have that

$$\begin{aligned} \mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta, \tau_{\mathcal{C}^\delta}^\eta < N \right) &\leq \mathbb{P}_x \left(\text{dist}_N \left(x^\eta, \Gamma_N^{\{x\}} \left(\tilde{\mathcal{B}}_{i,j} - 3\varepsilon \right) \right) > \frac{\delta'}{2} \right) \\ &\leq \exp \left(-\frac{\tilde{\mathcal{B}}_{i,j} - 4\varepsilon}{\eta} \right), \end{aligned} \quad (\text{A.3.131})$$

by [Corollary A.2.9](#).

Combining this bound with [\(A.3.129\)](#) yields

$$\mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta \right) \leq \exp \left(-\frac{\tilde{\mathcal{B}}_{i,j} - 4\varepsilon}{\eta} \right) + \exp \left(\frac{-\tilde{\mathcal{B}}_{i,j}}{\eta} \right), \quad (\text{A.3.132})$$

which concludes the proof. $\blacksquare$

Lemma A.3.26. *For any $\varepsilon > 0$, for any neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{K}_i$, $i = 1, \dots, K$ small enough, there exists $N \geq 0$, $\eta_0 > 0$ such that for all i, j , $x \in \mathcal{C}_i^\delta$, $0 < \eta < \eta_0$,*

$$\mathbb{Q}_{\mathcal{C}^\delta}^N(x, \mathcal{C}_j^\delta) \geq \exp \left(-\frac{\tilde{\mathcal{B}}_{i,j}}{\eta} - \frac{\varepsilon}{\eta} \right). \quad (\text{A.3.133})$$

Proof. For any i, j , there exists $N_{i,j} \geq 1$, $\xi^{i,j} \in (\mathbb{R}^d)^{N_{i,j}}$ such that $\xi_0^{i,j} \in \mathcal{K}_i$, $\xi_{N_{i,j}-1}^{i,j} \in \mathcal{K}_j$, $\xi_n^{i,j} \notin \bigcup_{l \neq i,j} \mathcal{K}_l$ for all $n = 1, \dots, N_{i,j} - 2$ and $\mathcal{S}_{N_{i,j}}(\xi^{i,j}) \leq \tilde{\mathcal{B}}_{i,j} + \varepsilon$. Define $\delta_{i,j} := \min \left\{ d(\xi_n^{i,j}, \bigcup_{l \neq i,j} \mathcal{K}_l) : n = 1, \dots, N_{i,j} - 2 \right\}$ and $\delta := \min_{i,j \in I} \delta_{i,j}$. By construction, it holds that $\delta > 0$.

Require that $\mathcal{C}_i^\delta$ be contained in $\mathcal{W}_i \cap \mathcal{U}_{\delta/2}(\mathcal{K}_i)$ for all $i = 1, \dots, K$. Now, given such $\mathcal{C}_i^\delta$ neighborhoods of $\mathcal{K}_i$, $i = 1, \dots, K$, there exists $0 < \delta' \leq \delta/2$ such that $\mathcal{U}'_\delta(\mathcal{K}_i)$ is contained in $\mathcal{C}_i^\delta$ for all $i = 1, \dots, K$.

Apply [Lemma A.3.10](#) to $\mathcal{K}_i$, $i = 1, \dots, K$ with $\varepsilon \leftarrow \min(\varepsilon, \delta'/2)$ and denote by N_i the bound on the length of paths obtained. Define

$$N := \max_{i \in I} N_i + 1. \quad (\text{A.3.134})$$

Fix $i, j \in I$ and $x \in \mathcal{C}_i^\delta$. Since $\mathcal{C}_i^\delta \subset \mathcal{W}_i$, there exists $z \in \mathcal{K}_i$ such that $\rho(x, z) < \varepsilon$. Moreover, note that $\rho(z, z) = 0$ since z is a critical point of f .

By [Lemma A.3.10](#), there exists $n \leq N$, $\xi \in (\mathbb{R}^d)^n$ such that $\xi_0 = z$, $\xi_{N-1} = \xi_0^{i,j}$, $\xi_k \in \mathcal{U}'_{\delta'/2}(\mathcal{K}_i)$ for all $k = 1, \dots, n-2$ and $\mathcal{S}_n(\xi) < \varepsilon$.

Considering the concatenation

$$\zeta := \left(\underbrace{x, z, z, \dots, z}_{N-n \text{ times}}, \xi_0, \xi_1, \dots, \xi_{n-2}, \xi_{n-1}, \xi_1^{i,j}, \dots, \xi_{N_{i,j}-1}^{i,j} \right) \quad (\text{A.3.135})$$

which is a path of length $N + N_{i,j}$ made of $x \in \mathcal{C}_i^\delta$, then exactly N points in $\mathcal{U}'_{\delta'/2}(\mathcal{K}_i)$ then $N_{i,j} - 2$ in $\mathbb{R}^d \setminus \mathcal{U}_{\delta/2}(\mathcal{C}^\delta)$ and $\xi_{N_{i,j}-1}^{i,j} \in \mathcal{K}_j$. Moreover, by construction, $\mathcal{S}_{N+N_{i,j}}(\zeta) \leq \tilde{\mathcal{B}}_{i,j} + 3\varepsilon$. Therefore, if

$$\text{dist}_{N+N_{i,j}}(x^\eta, \zeta) < \delta'/2, \quad (\text{A.3.136})$$

with $x_0^\eta = x$, then $x_1^\eta, \dots, x_N^\eta$ are in $\mathcal{U}'_{\delta'}(\mathcal{K}_i) \subset \mathcal{C}_i^\delta$ and, since $\delta' \leq \delta/2$, $x_{N+1}^\eta, \dots, x_{N+N_{i,j}-2}^\eta$ are not in $\mathcal{U}_{\delta/4}(\mathcal{C}^\delta)$, and therefore not in $\mathcal{C}^\delta$. Moreover, $x_{N+N_{i,j}-1}^\eta$ would be in $\mathcal{U}'_{\delta'/2}(\mathcal{K}_j) \subset \mathcal{C}_j^\delta$.

Thus, all the paths x^η satisfying (A.3.136) with $x_0^\eta = x$ are started at x and their N -th point that fall into $\mathcal{C}^\delta$ belongs to $\mathcal{C}_j^\delta$.

Therefore, using the definition of $\mathbb{Q}_{\mathcal{C}^\delta}^N$, we have that

$$\begin{aligned} \mathbb{Q}_{\mathcal{C}^\delta}^N(x, \mathcal{C}_j^\delta) &\geq \mathbb{P}_x \left(\text{dist}_{N+N_{i,j}}(x^\eta, \zeta) < \delta'/2 \right) \\ &\geq \exp \left(-\frac{\tilde{\mathcal{B}}_{i,j} + 4\varepsilon}{\eta} \right), \end{aligned} \quad (\text{A.3.137})$$

by [Corollary A.2.9](#). ■

Lemma A.3.27. *For any $i, j \in I$,*

$$\begin{aligned} \mathcal{B}_{i,j} &= \min \left\{ \sum_{l=0}^{n-2} \tilde{\mathcal{B}}_{i_l, i_{l+1}} : i_0 = i, i_{n-1} = j, i_l \in I \text{ for } l = 1, \dots, n-2, n \geq 1 \right\} \\ &= \min \left\{ \sum_{l=0}^{K-2} \tilde{\mathcal{B}}_{i_l, i_{l+1}} : i_0 = i, i_{K-1} = j, i_l \in I \text{ for } l = 1, \dots, K-2 \right\}. \end{aligned} \quad (\text{A.3.138})$$

Proof. It suffices to show that

$$\mathcal{B}_{i,j} = \min \left\{ \sum_{l=0}^{n-2} \widetilde{\mathcal{B}}_{i_l, i_{l+1}} : i_0 = i, i_{n-1} = j, i_l \in I \text{ for } l = 1, \dots, n-2, n \geq 1 \right\}. \quad (\text{A.3.139})$$

The statement of the lemma then follows from the fact that $\widetilde{\mathcal{B}}_{i,l} = 0$ for all $l \in I$ and that shortest paths on graphs can be chosen not to visit the same node twice.

For the inequality $(\geq)$, notice that any path between $\mathcal{K}_i$ and $\mathcal{K}_j$ can be decomposed into a concatenation of paths between $\mathcal{K}_i$ and $\mathcal{K}_{i_1}$, $\mathcal{K}_{i_1}$ and $\mathcal{K}_{i_2}$, $\dots$, $\mathcal{K}_{i_{n-1}}$ and $\mathcal{K}_j$ for some $i_1, \dots, i_{n-1} \in I$ that do not enter any other equivalence class in between. Therefore, the inequality $(\geq)$ follows from the definition of $\mathcal{B}_{i,j}$ and $\widetilde{\mathcal{B}}_{i_l, i_{l+1}}$.

We now focus on $(\leq)$.

Fix ε . Take $n \geq 1$, $i_0 = i$, $i_{n-1} = j$, $i_l \in I$ for $l = 1, \dots, n-2$. There are paths $\xi^0, \dots, \xi^{n-2}$ of lengths $N_0, \dots, N_{n-2}$ such that $\xi_0^l \in \mathcal{K}_{i_l}$, $\xi_{N_l-1}^l \in \mathcal{K}_{i_{l+1}}$, $S_{N_l}(\xi^l) \leq \widetilde{\mathcal{B}}_{i_l, i_{l+1}} + \varepsilon/n$ for $l = 0, \dots, n-2$.

By Lemma A.3.10, for all $l = 0, \dots, n-2$, $\xi_{N_l-1}^l$ and ξ_0^{l+1} can be connected by path of cost at most ε/n . Therefore concatenating all these paths yield ζ of length N with $\zeta_0 \in \mathcal{K}_i$, $\zeta_{N-1} \in \mathcal{K}_j$ and $S_N(\zeta) \leq \sum_{l=0}^{n-2} \widetilde{\mathcal{B}}_{i_l, i_{l+1}} + 2\varepsilon$. Since $S_N(\zeta) \geq \mathcal{B}_{i,j}$, we obtain the desired result. ■

Notation 1. We will write, for non-decreasing $f: \mathbb{R} \rightarrow \mathbb{R}$,

$$a \asymp f(b \pm c) \iff f(b - c) \leq a \leq f(b + c). \quad (\text{A.3.140})$$

Proposition A.3.28. For any $\varepsilon > 0$ and any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{K}_i$, $i = 1, \dots, K$, there exists $N \geq 0$, $\eta_0 > 0$ such that for all i, j , $x \in \mathcal{C}_i^\delta$, $0 < \eta < \eta_0$,

$$\mathbb{Q}_{\mathcal{C}^\delta}^N(x, \mathcal{C}_j^\delta) \asymp \exp\left(-\frac{\mathcal{B}_{i,j}}{\eta} \pm \frac{\varepsilon}{\eta}\right). \quad (\text{A.3.141})$$

Proof. Let us first start with $(\geq)$.

Let N satisfy the conditions of Lemma A.3.26 and define $N' := (K-1)N$. For any $i, j \in I$, by Lemma A.3.27, there exist $i_0 = i$, $i_{K-1} = j$, $i_l \in I$ for $l = 1, \dots, K-2$ such that

$$\mathcal{B}_{i,j} = \sum_{l=0}^{K-2} \widetilde{\mathcal{B}}_{i_l, i_{l+1}}. \quad (\text{A.3.142})$$

Therefore, we have that the probability of reaching $\mathcal{C}_j^\delta$ from $x \in \mathcal{C}_i^\delta$ in $N' = (K-1)N$ steps is greater than the probability of reaching sequentially the $\mathcal{C}_{i_{l+1}}^\delta$ from any point $x \in \mathcal{C}_{i_l}^\delta$, i.e.,

$$\begin{aligned} \inf_{x \in \mathcal{C}_i^\delta} \mathbb{Q}_{\mathcal{C}^\delta}^{N'}(x, \mathcal{C}_j^\delta) &\geq \prod_{l=0}^{K-2} \inf_{x \in \mathcal{C}_{i_l}^\delta} \mathbb{Q}_{\mathcal{C}^\delta}^N(x, \mathcal{C}_{i_{l+1}}^\delta) \\ &\geq \prod_{l=0}^{K-2} \exp\left(-\frac{\widetilde{\mathcal{B}}_{i_l, i_{l+1}}}{\eta} - \frac{\varepsilon}{\eta}\right) = \exp\left(-\frac{\mathcal{B}_{i,j}}{\eta} - \frac{(K-1)\varepsilon}{\eta}\right) \end{aligned} \quad (\text{A.3.143})$$

where we used Lemma A.3.26 to get the second inequality.

Now for the reverse inequality ($\leq$). Take $i, j \in I$ and denote $i_0 = i, i_{N'-1} = j$. By Lemma A.3.25, we have that

$$\begin{aligned} \sup_{x \in \mathcal{C}_i^\delta} \mathbb{Q}_{\mathcal{C}^\delta}^{N'}(x, \mathcal{C}_j^\delta) &\leq \sum_{i_1, \dots, i_{K-2} \in I} \prod_{l=0}^{K-2} \sup_{x \in \mathcal{C}_{i_l}^\delta} \mathbb{Q}_{\mathcal{C}^\delta}^N(x, \mathcal{C}_{i_{l+1}}^\delta) \\ &\leq \sum_{i_1, \dots, i_{K-2} \in I} \prod_{l=0}^{K-2} \exp\left(-\frac{\tilde{\mathcal{B}}_{i_l, i_{l+1}}}{\eta} + \frac{\varepsilon}{\eta}\right) \\ &\leq \sum_{i_1, \dots, i_{K-2} \in I} \exp\left(-\frac{\sum_{l=0}^{N'} \tilde{\mathcal{B}}_{i_l, i_{l+1}}}{\eta} + \frac{(K-1)\varepsilon}{\eta}\right). \end{aligned} \quad (\text{A.3.144})$$

Using Lemma A.3.27, we obtain that

$$\begin{aligned} \sup_{x \in \mathcal{C}_i^\delta} \mathbb{Q}_{\mathcal{C}^\delta}^{N'}(x, \mathcal{C}_j^\delta) &\leq \sum_{i_1, \dots, i_{K-2} \in I} \exp\left(-\frac{\mathcal{B}_{i,j}}{\eta} + \frac{(K-1)\varepsilon}{\eta}\right) \\ &= K^{K-2} \exp\left(-\frac{\mathcal{B}_{i,j}}{\eta} + \frac{(K-1)\varepsilon}{\eta}\right) \end{aligned} \quad (\text{A.3.145})$$

which concludes the proof. $\blacksquare$

Let us now give our first estimates on the invariant measure.

Definition A.7 ((Freidlin and Wentzell, 1998b, Chap. 6, §4)). Define, for every $i \in I$,

$$E_i = E(\mathcal{K}_i) := \min_{t \in \mathcal{T}_i} \sum_{(j \rightarrow l) \in t} \mathcal{B}_{j,l}, \quad (\text{A.3.146})$$

where $\mathcal{T}_i$ denotes the set of trees rooted at i in the complete graph on I .

Proposition A.3.29. For any $\mu_{\mathcal{C}^\delta}^\eta$ invariant probability measure for $\mathbb{Q}_{\mathcal{C}^\delta}^\delta$, the induced chain on $\mathcal{C}^\delta$, in the setting of Proposition A.3.28, for any $i \in I$,

$$\mu_{\mathcal{C}^\delta}^\eta(\mathcal{C}_i^\delta) \asymp \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j \in I} E(\mathcal{K}_j)}{\eta} \pm \frac{\varepsilon}{\eta}\right). \quad (\text{A.3.147})$$

Proof. If $\mu_{\mathcal{C}^\delta}^\eta$ is an invariant measure of $\mathbb{Q}_{\mathcal{C}^\delta}^\delta$, then it is an invariant measure of $\mathbb{Q}_{\mathcal{C}^\delta}^N$ for any N given by Proposition A.3.28.

Freidlin and Wentzell (1998b, Chap. 6, Lem. 3.1-3.2) combined with Proposition A.3.28 then give

$$\begin{aligned} &\exp(-(2 \text{ card } I + 2)\varepsilon/\eta) \frac{\exp(-E(\mathcal{K}_i)/\eta)}{\sum_{j \in I} \exp(-E(\mathcal{K}_j)/\eta)} \\ &\leq \mu_{\mathcal{C}^\delta}^\eta(\mathcal{C}_i^\delta) \\ &\leq \exp((2 \text{ card } I + 2)\varepsilon/\eta) \frac{\exp(-E(\mathcal{K}_i)/\eta)}{\sum_{j \in I} \exp(-E(\mathcal{K}_j)/\eta)}. \end{aligned} \quad (\text{A.3.148})$$

For η small enough, it holds that

$$\sum_{j \in I} \exp(-E(\mathcal{K}_j)/\eta) \asymp \exp\left(-\min_{j \in I} E(\mathcal{K}_j)/\eta \pm \varepsilon/\eta\right), \quad (\text{A.3.149})$$

which concludes the proof. $\blacksquare$

We now state a result that links invariant measures of $(x_n^\eta)_n$ and $\mathbb{Q}_C^\delta$. It is a consequence of Douc et al. (2018, Thm. 3.6.5).

Lemma A.3.30. *There is $\eta_0 > 0$ such that, for $0 < \eta \leq \eta_0$, if $(x_n^\eta)_n$ has an invariant probability measure μ_∞^η , then, for any $\mathcal{C}^\delta$ measurable neighborhood of crit f , we have that $\mu_\infty^\eta(\mathcal{C}^\delta) > 0$ and $\mu_{\mathcal{C}^\delta}^\eta$, the restriction of $\mu_\infty^\eta / \mu_\infty^\eta(\mathcal{C}^\delta)$ to $\mathcal{C}^\delta$ is an invariant measure for the induced chain on $\mathcal{C}^\delta$ and, for any measurable set $H \subset \mathbb{R}^d$,*

$$\frac{\mu_\infty^\eta(H)}{\mu_\infty^\eta(\mathcal{C}^\delta)} = \int_{\mathcal{C}^\delta} d\mu_{\mathcal{C}^\delta}^\eta(x) \mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in H\} \right]. \quad (\text{A.3.150})$$

Proof. We invoke the first item of Lemma A.3.22: there exists $\mathcal{X} \subset \mathbb{R}^d$ a compact set, $\eta_0 > 0$, such that for any $\mathcal{X}' \subset \mathbb{R}^d$ compact set such that $\mathcal{X} \subset \mathcal{X}'$, there exists $\alpha_0 > 0$ such that,

$$\forall \eta \leq \eta_0, x \in \mathcal{X}', \quad \mathbb{E}_x \left[e^{\frac{\alpha_0 \tau_{\mathcal{X}}^\eta}{\eta}} \right] < +\infty. \quad (\text{A.3.151})$$

Without loss of generality, at the potential expense of expanding $\mathcal{X}$, assume that $\mathcal{C}^\delta \subset \mathcal{X}$.

By applying Lemma A.3.24 to $\mathcal{U} \leftarrow \mathcal{C}^\delta$ and $\mathcal{X} \leftarrow \mathcal{X}$, we get that there is some $\eta_0, \alpha_0, > 0$ such that, for any $\eta \leq \eta_0, x \in \mathcal{X}$,

$$\mathbb{E}_x \left[e^{\frac{\alpha \tau_{\mathcal{C}^\delta}^\eta}{\eta}} \right] < +\infty. \quad (\text{A.3.152})$$

In particular, we have that $\mathbb{P}_x(\tau_{\mathcal{C}^\delta}^\eta < \infty) = 1$ for any $x \in \mathcal{X}$, and *a fortiori* for any $x \in \mathcal{C}^\delta$.

Let us now show that, for any $x \in \mathbb{R}^d$, $\mathbb{P}_x(\tau_{\mathcal{C}^\delta}^\eta < \infty) = 1$. Fix $x \in \mathbb{R}^d$. By choosing $\mathcal{X}'$ large enough to contain both $\mathcal{X}$ and x , (A.3.151) implies that, with $x_0 = x$, $\tau_{\mathcal{X}}^\eta < \infty$ almost surely. Therefore, by the strong Markov property, it holds that,

$$\mathbb{P}_x(\tau_{\mathcal{C}^\delta}^\eta < \infty) \geq \inf_{z \in \mathcal{X}} \mathbb{P}_z(\tau_{\mathcal{C}^\delta}^\eta < \infty) = 1. \quad (\text{A.3.153})$$

Therefore the assumptions of Douc et al. (2018, Thm. 3.6.5) are satisfied as well as its first item which yields the result. $\blacksquare$

We reach the next proposition, which is the first part of our main result. It is an adaptation of Freidlin and Wentzell (1998b, Thm. 4.1) to the discrete time setting.

Proposition A.3.31. *For any $\varepsilon > 0$, for any $\mathcal{C}_1^\delta, \dots, \mathcal{C}_K^\delta$ measurable neighborhoods of $\mathcal{K}_1, \dots, \mathcal{K}_K$ small enough, there exists $\eta_0 > 0$ such that for any $0 < \eta < \eta_0$, μ_∞^η invariant probability measure for $(x_n^\eta)_n$, for any $i \in I$,*

$$\mu_\infty^\eta(\mathcal{C}_i^\delta) \asymp \exp \left(-\frac{E(\mathcal{K}_i) - \min_{j \in I} E(\mathcal{K}_j)}{\eta} \pm \frac{\varepsilon}{\eta} \right). \quad (\text{A.3.154})$$

Proof. Let us first provide estimates for the unnormalized measure defined by, for any measurable set H ,

$$\tilde{\mu}(H) := \int_H d\mu_{\mathcal{C}^\delta}^\eta(x) \mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in H\} \right]. \quad (\text{A.3.155})$$

By definition of $\tau_{\mathcal{C}^\delta}^\eta$, in the sequence of points $x_0^\eta, \dots, x_{\tau_{\mathcal{C}^\delta}^\eta - 1}^\eta$, only x_0^η can be in $\mathcal{C}^\delta$. Therefore, for any $i, j \in I$, $x \in \mathcal{C}_i^\delta$,

$$\mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in \mathcal{C}_i^\delta\} \right] = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A.3.156})$$

Thus, we have that,

$$\begin{aligned} \tilde{\mu}(\mathcal{C}_i^\delta) &= \mu_{\mathcal{C}^\delta}^\eta(\mathcal{C}_i^\delta) \\ &\asymp \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j \in I} E(\mathcal{K}_j)}{\eta} \pm \frac{\varepsilon}{\eta}\right), \end{aligned} \quad (\text{A.3.157})$$

by [Proposition A.3.29](#). It now remains to estimate the normalization constant $\tilde{\mu}(\mathbb{R}^d)$. On the one hand, we have that

$$\begin{aligned} \tilde{\mu}(\mathbb{R}^d) &\geq \max_{i \in I} \tilde{\mu}(\mathcal{C}_i^\delta) \\ &\geq \max_{i \in I} \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j \in I} E(\mathcal{K}_j)}{\eta} - \frac{\varepsilon}{\eta}\right) = \exp\left(-\frac{\varepsilon}{\eta}\right). \end{aligned} \quad (\text{A.3.158})$$

On the other hand, by [Lemma A.3.24](#) applied with $\mathcal{U} \leftarrow \mathcal{C}^\delta$ and $\mathcal{X} \leftarrow \text{cl } \mathcal{C}^\delta$ (choosing $\mathcal{C}^\delta$ small enough so that is bounded), the quantity

$$c := \sup\{\mathbb{E}_x[\tau_{\mathcal{C}^\delta}^\eta] : x \in \mathcal{C}^\delta, 0 < \eta \leq \eta_0\} \quad (\text{A.3.159})$$

is finite. Therefore, we have that

$$\begin{aligned} \tilde{\mu}(\mathbb{R}^d) &= \int_{\mathbb{R}^d} d\mu_{\mathcal{C}^\delta}^\eta(x) \mathbb{E}_x[\tau_{\mathcal{C}^\delta}^\eta] \\ &\leq c \int_{\mathbb{R}^d} d\mu_{\mathcal{C}^\delta}^\eta(x) = c, \end{aligned} \quad (\text{A.3.160})$$

which, along with choosing η_0 small enough so that $c \leq \exp(\varepsilon/\eta)$, concludes the proof. $\blacksquare$

We will need the following lemma to prove the second part of our main result.

Lemma A.3.32. *For any $\varepsilon > 0$, for any $\mathcal{C}_1^\delta, \dots, \mathcal{C}_K^\delta$ measurable neighborhoods of $\mathcal{K}_1, \dots, \mathcal{K}_K$ small enough, D measurable set, $\delta_D > 0$, there exists $\eta_0 > 0$ such that, for any $0 < \eta < \eta_0$, for any $i \in I$, $x \in \mathcal{C}_i^\delta$,*

$$\mathbb{P}_x\left(\tau_{D-\delta_D}^\eta < \tau_{\mathcal{C}^\delta}^\eta\right) \leq \exp\left(-\frac{B(\mathcal{K}_i, D) - \varepsilon}{\eta}\right), \quad (\text{A.3.161})$$

where

$$D_{-\delta_D} := \{z \in D : d(z, \mathbb{R}^d \setminus D) \geq \delta_D\}. \quad (\text{A.3.162})$$

Proof. The requirement on the $\mathcal{C}_i^\delta$ are the same in the proof of [Lemma A.3.25](#) but we restate them here for completeness.

Assume that, without loss of generality, ε is small enough so that [Lemma A.3.9](#) with $r \leftarrow \varepsilon$ can be applied to every $\mathcal{K}_i$, $i = 1, \dots, K$. Denote by $\mathcal{W}_i$, $i = 1, \dots, K$ the corresponding neighborhoods of $\mathcal{K}_i$. Since these $\mathcal{W}_i$'s are neighborhoods of the $\mathcal{K}_i$'s, there exists $\delta > 0$ such that $\mathcal{U}_{2\delta}(\mathcal{K}_i) \subset \mathcal{W}_i$ for all $i = 1, \dots, K$. Require then that $\mathcal{C}_i^\delta$ be contained in $\mathcal{U}_\delta(\mathcal{K}_i)$ so that $\mathcal{U}_\delta(\text{cl } \mathcal{C}_i^\delta) \subset \mathcal{W}_i$ for all $i = 1, \dots, K$. Moreover, assume that $\delta > 0$ is small enough so that the neighborhoods $\mathcal{U}_\delta(\mathcal{K}_i)$, $i = 1, \dots, K$ are pairwise disjoint and that $\delta \leq \delta_D$. Define $0 < \delta' \leq \delta$ such that $\mathcal{U}'_\delta(\mathcal{K}_i)$ is contained in $\mathcal{C}_i^\delta$ for all $i = 1, \dots, K$.

Fix $i \in I$ and consider $\xi \in (\mathbb{R}^d)^N$ such that $\xi_0 \in \mathcal{U}'_\delta(\text{cl } \mathcal{C}_i^\delta)$, $\xi_{N-1} \in \mathcal{U}'_\delta(D_{-\delta_D})$. By construction, $\xi_0 \in \mathcal{U}_\delta(\mathcal{C}_i^\delta) \subset \mathcal{W}_i$. Therefore, there exists $x \in \mathcal{K}_i$ such that $\rho(x, \xi_0) < \varepsilon$. Moreover, since $\delta' \leq \delta \leq \delta_D$, $\mathcal{U}'_\delta(D_{-\delta_D}) \subset D$ so that $\xi_{N-1} \in D$.

Define $\zeta := (x, \xi_0, \dots, \xi_{N-1})$ which is a path from $\mathcal{K}_i$ to D so that

$$\mathcal{S}_N(\xi) \geq \mathcal{S}_{N+1}(\zeta) - \varepsilon \geq \mathcal{B}(\mathcal{K}_i, D) - \varepsilon. \quad (\text{A.3.163})$$

We now follow the same outline as for the proof of [Lemma A.3.25](#). Fix $x \in \mathcal{C}_i^\delta$. For any $N \geq 0$, we have that

$$\mathbb{P}_x\left(\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta\right) \leq \mathbb{P}_x\left(\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta, \tau_{\mathcal{C}^\delta}^\eta < N\right) + \mathbb{P}_x\left(\tau_{\mathcal{C}^\delta}^\eta \geq N\right). \quad (\text{A.3.164})$$

For some N large enough, [Lemma A.3.24](#) yields

$$\mathbb{P}_x\left(\tau_{\mathcal{C}^\delta}^\eta \geq N\right) \leq \exp\left(-\frac{\mathcal{B}(\mathcal{K}_i, D) - \varepsilon}{\eta}\right). \quad (\text{A.3.165})$$

We now bound $\mathbb{P}_x\left(\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta, \tau_{\mathcal{C}^\delta}^\eta < N\right)$ for this choice of N . For this, it suffices to note that $\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta, \tau_{\mathcal{C}^\delta}^\eta < N$ implies that $\text{dist}_N\left(x^\eta, \Gamma_N^{\{x\}}(\xi - 2\varepsilon)\right) > \frac{\delta'}{2}$. Then, applying [Corollary A.2.9](#) yields

$$\mathbb{P}_x\left(\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta, \tau_{\mathcal{C}^\delta}^\eta < N\right) \leq \exp\left(-\frac{\mathcal{B}(\mathcal{K}_i, D) - 3\varepsilon}{\eta}\right), \quad (\text{A.3.166})$$

which concludes the proof. $\blacksquare$

We can use this lemma to upper-bound the $\mu_\infty^\eta(D)$.

Lemma A.3.33. *For any $\varepsilon > 0$, any bounded measurable set D , there exists $\eta_0 > 0$ such that for any $0 < \eta < \eta_0$, μ_∞^η invariant probability measure for $(x_n^\eta)_n$, for any $i \in I$,*

$$\mu_\infty^\eta(D_{-\delta_D}) \leq \exp\left(-\frac{\min_{i \in I}\{E(\mathcal{K}_i) + \mathcal{B}(\mathcal{K}_i, D)\} - \min_{i \in I} E(\mathcal{K}_i)}{\eta} + \frac{\varepsilon}{\eta}\right), \quad (\text{A.3.167})$$

where

$$D_{-\delta_D} := \{z \in \mathbb{R}^d : d(z, \mathbb{R}^d \setminus D) \geq \delta_D\}. \quad (\text{A.3.168})$$

Proof. Using $\mathcal{C}_1^\delta, \dots, \mathcal{C}_K^\delta$ measurable neighborhoods of $\mathcal{K}_1, \dots, \mathcal{K}_K$ small enough given by [Lemma A.3.32](#), we provide an estimate for the weight of D for the unnormalized measure $\tilde{\mu}$ defined in the proof of [Proposition A.3.31](#), i.e., for

$$\tilde{\mu}(D_{-\delta_D}) := \int_{\mathcal{C}^\delta} d\mu_{\mathcal{C}^\delta}^\eta(x) \mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\} \right], \quad (\text{A.3.169})$$

and the result will follow from the estimate on the normalization constant $\tilde{\mu}(\mathbb{R}^d)$ obtained in the proof of [Proposition A.3.31](#).

By [Lemma A.3.24](#) applied with $\mathcal{U} \leftarrow \mathcal{C}^\delta$ and $\mathcal{X} \leftarrow \text{cl } \mathcal{C}^\delta \cup D$ (choosing $\mathcal{C}^\delta$ small enough so that is bounded), the quantity

$$c := \sup\{\mathbb{E}_x[\tau_{\mathcal{C}^\delta}^\eta] : x \in \mathcal{C}^\delta \cup D, 0 < \eta \leq \eta_0\} \quad (\text{A.3.170})$$

is finite.

Fix $i \in I$ and $x \in \mathcal{C}_i^\delta$. If $\tau_{D_{-\delta_D}}^\eta \geq \tau_{\mathcal{C}^\delta}^\eta$, then $\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\}$ would be 0 so, we have that

$$\begin{aligned} \mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\} \right] &= \mathbb{E}_x \left[\mathbb{1}\{\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta\} \sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\} \right] \\ &\leq \mathbb{E}_x \left[\mathbb{1}\{\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta\} \tau_{\mathcal{C}^\delta}^\eta \right] \\ &= \mathbb{E}_x \left[\mathbb{1}\{\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta\} \left(\tau_{D_{-\delta_D}}^\eta + \mathbb{E}_{x_{\tau_{D_{-\delta_D}}^\eta}^\eta} \left[\tau_{\mathcal{C}^\delta}^\eta \right] \right) \right], \end{aligned} \quad (\text{A.3.171})$$

where we used the strong Markov property. Bounding $\tau_{D_{-\delta_D}}^\eta$ by $\tau_{\mathcal{C}^\delta}^\eta$ on the event $\{\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta\}$, we obtain that

$$\begin{aligned} \mathbb{E}_x \left[\sum_{n=0}^{\tau_{\mathcal{C}^\delta}^\eta - 1} \mathbb{1}\{x_n^\eta \in D_{-\delta_D}\} \right] &\leq \mathbb{E}_x \left[\mathbb{1}\{\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta\} \left(\tau_{\mathcal{C}^\delta}^\eta + \mathbb{E}_{x_{\tau_{D_{-\delta_D}}^\eta}^\eta} \left[\tau_{\mathcal{C}^\delta}^\eta \right] \right) \right] \\ &\leq 2c \mathbb{P}_x \left(\tau_{D_{-\delta_D}}^\eta < \tau_{\mathcal{C}^\delta}^\eta \right) \\ &\leq 2c \exp \left(-\frac{\mathcal{B}(\mathcal{K}_i, D) - \varepsilon}{\eta} \right), \end{aligned} \quad (\text{A.3.172})$$

where we invoked [Lemma A.3.32](#) for the last inequality.

Combining this bound and [Proposition A.3.29](#), we obtain that

$$\begin{aligned} \tilde{\mu}(D_{-\delta_D}) &\leq 2c \sum_{i \in I} \mu_{\mathcal{C}^\delta}^\eta(\mathcal{C}_i^\delta) \exp \left(-\frac{\mathcal{B}(\mathcal{K}_i, D) - \varepsilon}{\eta} \right) \\ &\leq 2c \text{card } I \exp \left(-\frac{\min_{i \in I} \{E(\mathcal{K}_i) + \mathcal{B}(\mathcal{K}_i, D)\} - \min_{i \in I} E(\mathcal{K}_i)}{\eta} + \frac{2\varepsilon}{\eta} \right), \end{aligned} \quad (\text{A.3.173})$$

which concludes the proof. $\blacksquare$

A.3.6 Convergence and stability

Let us first begin by showing that, for every initial point, the flow of f converges to one of the $\mathcal{K}_i$.

Lemma A.3.34. *For any $x \in \mathbb{R}^d$, there is $i \in I$ such that*

$$\lim_{t \rightarrow +\infty} d(\Theta_t(x), \mathcal{K}_i) = 0. \quad (\text{A.3.174})$$

Proof. Fix $x \in \mathbb{R}^d$. For any $t \geq 0$, $\Theta_t(x)$ belongs to $\{z \in \mathbb{R}^d : f(z) \leq f(x)\}$, which is compact by coercivity of f . Therefore, the set of accumulation points of $(\Theta_t(x))_{t \geq 0}$ is non-empty, connected and included in $\text{crit } f$. Therefore, since the $\mathcal{K}_i$, for $i \in I$, are the connected components of $\text{crit } f$, there is $i \in I$ such that the accumulation points of $(\Theta_t(x))_{t \geq 0}$ all belong to $\mathcal{K}_i$.

If $d(\Theta_t(x), \mathcal{K}_i)$ did not converge to 0, there would be a subsequence that would converge to some point out of any $\mathcal{K}_i$, which would be a contradiction. Therefore, $d(\Theta_t(x), \mathcal{K}_i)$ must converge to 0. ■

Let us restate the definition of minimizing component that we introduced in the main text.

Definition A.8. For any $i \in I$, we say that $\mathcal{K}_i$ is a *minimizing component* if there exists $\mathcal{U}$ a neighborhood of $\mathcal{K}_i$ such that,

$$\arg \min_{x \in \mathcal{U}} f(x) = \mathcal{K}_i. \quad (\text{A.3.175})$$

$\mathcal{K}_i$ is called non-minimizing otherwise.

We now state a standard definition for asymptotic stability.

Definition A.9. A connected component of the critical points $\mathcal{K}_i$, for some $i \in I$, is said to be *asymptotically stable* if there exists $\mathcal{U}$ a neighborhood of $\mathcal{K}_i$ such that, for any $x \in \mathcal{U}$, $\Theta_t(x)$ converges to $\mathcal{K}_i$.

The notions of minimizing component and asymptotic stability are equivalent in our context.

Lemma A.3.35. *For any $i \in I$, $\mathcal{K}_i$ is a minimizing component if and only if it is asymptotically stable.*

Proof. We start with the direct implication. Assume that $\mathcal{K}_i$ is a minimizing component. Therefore, there exists $\delta > 0$ such that

$$\arg \min_{x \in \text{cl } \mathcal{U}_\delta(\mathcal{K}_i)} f(x) = \mathcal{K}_i. \quad (\text{A.3.176})$$

Moreover, assume that δ is small enough so that, $\text{cl } \mathcal{U}_\delta(\mathcal{K}_i) \cap \text{crit } f = \mathcal{K}_i$.

Then, for any $x \in \mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{K}_i$, $f(x) > f(\mathcal{K}_i)$. By continuity of f , compactness of $\text{cl } \mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{U}_{\delta/2}(\mathcal{K}_i)$ and definition of minimizing component, we have that

$$\min_{x \in \text{cl } \mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{U}_{\delta/2}(\mathcal{K}_i)} f(x) > f(\mathcal{K}_i). \quad (\text{A.3.177})$$

Define $\delta' := \frac{1}{2} \min_{x \in \text{cl } \mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{U}_{\delta/2}(\mathcal{K}_i)} \{f(x) - f(\mathcal{K}_i)\}$ and

$$\mathcal{C}^\delta := \{x \in \mathcal{U}_{\delta/2}(\mathcal{K}_i) : f(x) < f(\mathcal{K}_i) + \delta'\} \subset \mathcal{U}_\delta(\mathcal{K}_i), \quad (\text{A.3.178})$$

which is a neighborhood of $\mathcal{K}_i$ by continuity of f .

We now show that trajectories of the flow starting in $\mathcal{C}^\delta$ converge to $\mathcal{K}_i$. Take $x \in \mathcal{C}^\delta$. Since $f(\Theta_t(x))$ is non-increasing, then $(\Theta_t(x))_t$ remains in $\{x \in \mathbb{R}^d : f(x) < f(\mathcal{K}_i) + \delta'\}$. $(\Theta_t(x))_t$ also cannot escape from $\mathcal{U}_{\delta/2}(\mathcal{K}_i)$: if it were, by continuity of the flow and of the distance to $\mathcal{K}_i$, there would be $t > 0$ such that $d(\Theta_t(x), \mathcal{K}_i) = \delta/2$, which is a contradiction since $f(\Theta_t(x)) < f(\mathcal{K}_i) + \delta'$. Hence, $(\Theta_t(x))_t$ remains in $\mathcal{U}_{\delta/2}(\mathcal{K}_i)$ and so in $\mathcal{C}^\delta$. By [Lemma A.3.34](#), $(\Theta_t(x))_t$ converges to some component of the critical points, which must be $\mathcal{K}_i$ since $\mathcal{C}^\delta$ is disjoint from the other components.

We now show the converse implication. Since $\mathcal{K}_i$ is a connected component of $\text{crit } f$, f is constant on $\mathcal{K}_i$. Denote by f_i^* this value. Take $\delta > 0$ small enough such that $\mathcal{U}_\delta(\mathcal{K}_i) \cap \text{crit } f = \mathcal{K}_i$ and such that, for any $x \in \mathcal{U}_\delta(\mathcal{K}_i)$, $(\Theta_t(x))_t$ converges to $\mathcal{K}_i$. Take $x \in \mathcal{U}_\delta(\mathcal{K}_i) \setminus \mathcal{K}_i$. We show that $f(x) > f_i^*$. For any $t > 0$,

$$f(\Theta_t(x)) - f(x) = - \int_0^t \|\nabla f(\Theta_s(x))\|^2 ds, \quad (\text{A.3.179})$$

which must (strictly) negative since x is not a critical point of f . Since $f(\Theta_t(x))$ is non-increasing in t and lower-bounded — because $\inf_{\mathbb{R}^d} f > -\infty$ by coercivity of f —, it must converge as $t \rightarrow +\infty$ and its limit satisfies $\lim_{t \rightarrow +\infty} f(\Theta_t(x)) < f(x)$. Moreover, $(\Theta_t(x))_t$ converges to $\mathcal{K}_i$ so that all its accumulation points belong to $\mathcal{K}_i$ and have the same objective value f_i^* . Hence, we have that $\lim_{t \rightarrow +\infty} f(\Theta_t(x)) = f_i^*$ and $f_i^* < f(x)$. ■

In the following, we will thus use the terms minimizing component and asymptotically stable interchangeably.

The next lemma shows that if $\mathcal{K}_i$ is not asymptotically stable, then it is unstable in the sense of [Freidlin and Wentzell \(1998b, Chap. 6, §4\)](#).

Lemma A.3.36. *If $\mathcal{K}_i$ is not asymptotically stable, then there exists $j \in I$ such that $\mathcal{B}_{i,j} = 0$.*

Proof. If $\mathcal{K}_i$ is not asymptotically stable, then, for every $n \geq 1$, there exists $x_n \in \mathcal{U}_{1/n}(\mathcal{K}_i)$ such that $\Theta_t(x_n)$ does not converge to $\mathcal{K}_i$. By [Lemma A.3.34](#), there exists $j_n \in I$ such that $(\Theta_t(x_n))_t$ converges to $\mathcal{K}_{j_n}$. Since I is finite, there exists $j \in I$ such that $j_n = j$ for infinitely many n . Replacing $(x_n)_{n \geq 1}$ by a subsequence, we can assume that $(\Theta_t(x_n))_t$ converges to $\mathcal{K}_j$ for all $n \geq 1$.

We now show that $\mathcal{B}_{i,j} = 0$.

Fix $\varepsilon > 0$ and consider $\mathcal{W}_\varepsilon(\mathcal{K}_i), \mathcal{W}_\varepsilon(\mathcal{K}_j)$ which are neighborhoods of $\mathcal{K}_i, \mathcal{K}_j$ by [Lemma A.3.9](#). Since $(x_n)_{n \geq 1}$ converges to $\mathcal{K}_i$, there exists n such that $x_n \in \mathcal{W}_\varepsilon(\mathcal{K}_i)$. In turn, since $\Theta_t(x_n)$ converges to $\mathcal{K}_j$, there exists $T > 0$, which we can choose integer, such that, $\Theta_T(x_n) \in \mathcal{W}_\varepsilon(\mathcal{K}_j)$. Therefore, there exists $x \in \mathcal{K}_i$ and $z \in \mathcal{K}_j$ such that $\rho(x, x_n) < \varepsilon$ and $\rho(\Theta_T(x_n), z) < \varepsilon$. Consider the discrete path $\xi \in (\mathbb{R}^d)^{T+3}$ defined by $\xi_0 = x, \xi_n = \Theta_{n-1}(x)$ for all $1 \leq n \leq T+1$ and $\xi_{T+2} = z$. By [Lemma A.3.2](#), we have that

$$\mathcal{S}_{T+3}(\xi) = \rho(x, x_n) + \rho(\Theta_T(x), z) < 2\varepsilon, \quad (\text{A.3.180})$$

and therefore, since $\xi_0 \in \mathcal{K}_i$ and $\xi_{T+2} \in \mathcal{K}_j$, $\mathcal{B}_{i,j} < 2\varepsilon$. ■

The following lemma is now a straightforward adaptation of [Freidlin and Wentzell \(1998b, Chap. 6, Lemma 4.2\)](#).

Lemma A.3.37. *If $\mathcal{K}_i$ is not asymptotically stable, then there exists $j \in I$ such that $\mathcal{B}_{i,j} = 0$ and such that, for any $l \in I$, $\mathcal{B}_{j,l} > 0$.*

Proof. For the sake of contradiction, assume that such a j does not exist. Then we build an infinite sequence $j_0 = i, j_1, \dots$ such that $\mathcal{B}_{j_n, j_{n+1}} = 0$ for all $n \geq 0$. By definition of equivalence classes, any j cannot appear twice in this sequence. But since I is finite, this is a contradiction. ■

We now reach our final result on unstable equivalence classes: they have negligible weight in the invariant measure.

Lemma A.3.38. *If $\mathcal{K}_i$ is not asymptotically stable, or, equivalently, non-minimizing, then there exists $j \in I$ such that $\mathcal{K}_j$ is asymptotically stable, $\mathcal{B}_{i,j} = 0$ and,*

$$E_j < E_i \quad (\text{A.3.181})$$

Proof. By Lemma A.3.37, there exists $j \in I$ such that $\mathcal{B}_{i,j} = 0$ and such that, for any $l \in I$, $\mathcal{B}_{j,l} > 0$. Lemma A.3.36 implies in particular that $\mathcal{K}_j$ must be asymptotically stable

It remains to show that $E_j < E_i$. Take $t \in \mathcal{T}_i$ such that

$$E_i = \sum_{(l \rightarrow k) \in t} \mathcal{B}_{l,k}. \quad (\text{A.3.182})$$

j has an outgoing edge in t and denote by j' the other end of that edge. Now, consider the tree $t' \in \mathcal{T}_j$ obtained from t by removing the outgoing edge from j to j' and adding an edge from i to j . Then, by definition of E_j ,

$$E_j \leq \sum_{(l \rightarrow k) \in t'} \mathcal{B}_{l,k} = \sum_{(l \rightarrow k) \in t} \mathcal{B}_{l,k} - \mathcal{B}_{j,j'} + \mathcal{B}_{i,j}. \quad (\text{A.3.183})$$

But, by definition of j , $\mathcal{B}_{i,j} = 0$ and $\mathcal{B}_{j,j'} > 0$. Therefore, $E_j < E_i$. ■

We now show the second part of our main result: the invariant measure concentrates on the ground states, which are asymptotically stable by Lemma A.3.38.

For this, we need the following lemma.

Lemma A.3.39. *For any $i \in I$ such that $\mathcal{K}_i$ is minimizing, there exists $\delta > 0$ such that, for any $0 < \delta' \leq \delta$,*

$$\mathcal{B}(\mathcal{K}_i, \mathbb{R}^d \setminus \mathcal{U}'_\delta(\mathcal{K}_i)) > 0. \quad (\text{A.3.184})$$

Proof. Since $\mathcal{K}_i$ is minimizing, there exists $\delta > 0$ such that, for any $x \in \mathcal{U}_\delta(\mathcal{K}_i)$, $f(x) > f_i$ where f_i is the value of f on $\mathcal{K}_i$. Take $\delta' \leq \delta$, $\mathcal{U} := \mathcal{U}'_\delta(\mathcal{K}_i)$ and

$$\Delta := \min\{V_\infty(x) - \alpha_\infty(f_i) : x \in \mathbb{R}^d, d(x, \mathcal{K}_i) = \delta'/2\}. \quad (\text{A.3.185})$$

Then, by the continuity of V_∞ and the fact that α_∞ is (strictly) increasing, we have that $\Delta > 0$. To conclude the proof of this lemma, we now show that $\mathcal{B}(\mathcal{K}_i, \mathbb{R}^d \setminus \mathcal{U}) \geq \frac{\Delta}{2}$. Consider some $T > 0$ and $\gamma \in \mathcal{C}([0, T], \mathbb{R}^d)$ such that $\gamma_0 \in \mathcal{K}_i$ and $\gamma_T \in \mathbb{R}^d \setminus \mathcal{W}$. By continuity of γ and $d(\cdot, \mathcal{K}_i)$, there exists $t \in [0, T]$ such that $d(\gamma_t, \mathcal{K}_i) = \delta'/2$. By the same computation as in Lemma A.3.7, we have that

$$\begin{aligned} \Delta &\leq V_\infty(\gamma_t) - V_\infty(\gamma_0) \\ &= 2 \int_0^t \frac{\langle \dot{\gamma}_s, \nabla f(\gamma_s) \rangle}{\sigma_\infty^2(f(\gamma_s))} ds \\ &\leq 2\mathcal{S}_{[0,t]}(\gamma) \\ &\leq 2\mathcal{S}_{[0,T]}(\gamma). \end{aligned} \quad (\text{A.3.186})$$

Since this is valid for any γ , we obtain that $\mathcal{B}(\mathcal{K}_i, \mathbb{R}^d \setminus \mathcal{U}) \geq \frac{\Delta}{2}$. ■

The next proposition shows that the invariant measure concentrates exponentially on states that are asymptotically stable (and contain the ground states).

Proposition A.3.40. *Consider $J \subset I$ such that, for all $i \in J$, $\mathcal{K}_i$ is minimizing and, such that, J contains $\arg \min_{i \in I} E_i$. Consider $\mathcal{C}_i^\delta$ small enough neighborhoods of $\mathcal{K}_i$ for $i \in J$. Then, there exists $c > 0$, $\eta_0 > 0$ such that, for any $\eta \leq \eta_0$, for any μ_∞^η invariant measure of $(x_n^\eta)_{n \geq 0}$,*

$$\mu_\infty^\eta \left(\mathbb{R}^d \setminus \bigcup_{i \in J} \mathcal{C}_i^\delta \right) \leq e^{-\frac{c}{\eta}}. \quad (\text{A.3.187})$$

Proof. Take $\delta > 0$ small enough so that, for any $i \in J$, $\mathcal{U}_\delta(\mathcal{K}_i) \subset \mathcal{C}_i^\delta$ and

$$d(\mathcal{K}_i, \mathbb{R}^d \setminus \mathcal{U}_\delta(\mathcal{K}_i)) > 0. \quad (\text{A.3.188})$$

This is possible by [Lemma A.3.39](#).

Define

$$D := \mathbb{R}^d \setminus \bigcup_{i \in J} \mathcal{U}_{\delta/2}(\mathcal{K}_i). \quad (\text{A.3.189})$$

With the notations of [Lemma A.3.33](#), we show that

$$\mathbb{R}^d \setminus \bigcup_{i \in J} \mathcal{C}_i^\delta \subset D_{-\delta/2}. \quad (\text{A.3.190})$$

Indeed, take $x \in \mathbb{R}^d \setminus \bigcup_{i \in J} \mathcal{C}_i^\delta$. Then, for any $i \in J$, it holds that $d(x, \mathcal{K}_i) \geq \delta$ and so, we have

$$d(x, \mathbb{R}^d \setminus D) = d\left(x, \mathbb{R}^d \setminus \bigcup_{i \in J} \mathcal{U}_{\delta/2}(\mathcal{K}_i)\right) \geq \frac{\delta}{2}. \quad (\text{A.3.191})$$

Hence $x \in D_{-\delta/2}$ and, it suffices to bound $\mu_\infty^\eta(D_{-\delta/2})$ to show the result.

Moreover, for any $i \in J$,

$$\mathcal{B}(\mathcal{K}_i, D) \geq \mathcal{B}(\mathcal{K}_i, \mathbb{R}^d \setminus \mathcal{U}_{\delta/2}(\mathcal{K}_i)) > 0, \quad (\text{A.3.192})$$

so that the quantity

$$c := \min \left(\min_{i \notin J} E_i - \min_{i \in I} E_i, \min_{i \in J} \mathcal{B}(\mathcal{K}_i, D) \right) \quad (\text{A.3.193})$$

is positive.

Apply [Lemma A.3.33](#) with $\delta_D \leftarrow \delta/2$, $\varepsilon \leftarrow c/2$ to get that, for any $\eta \leq \eta_0$,

$$\mu_\infty^\eta(D_{-\delta/2}) \leq \exp \left(-\frac{\min_{i \in I} \{E(\mathcal{K}_i) + \mathcal{B}(\mathcal{K}_i, D)\} - \min_{j \in I} E(\mathcal{K}_j)}{\eta} + \frac{c}{2\eta} \right). \quad (\text{A.3.194})$$

But, for any $i \in I$, the exponent can be estimated as

$$\{E(\mathcal{K}_i) + \mathcal{B}(\mathcal{K}_i, D)\} - \min_{j \in I} E(\mathcal{K}_j) \geq \begin{cases} E(\mathcal{K}_i) - \min_{j \in I} E(\mathcal{K}_j) & \text{if } i \notin J \\ \mathcal{B}(\mathcal{K}_i, D) & \text{if } i \in J, \end{cases} \quad (\text{A.3.195})$$

which is always positive, even in the first case, since $\arg \min_{j \in I} E(\mathcal{K}_j) \subset J$. Therefore, it holds that

$$\min_{i \in I} \{E(\mathcal{K}_i) + \mathcal{B}(\mathcal{K}_i, D)\} - \min_{j \in I} E(\mathcal{K}_j) - \varepsilon \geq c - \varepsilon = \frac{c}{2}, \quad (\text{A.3.196})$$

which concludes the proof. $\blacksquare$

A.3.7 Main results

In this section, we restate the main results [Theorem 4.3](#), [Theorem 4.4](#), [Theorem 4.6](#), and [Theorem 4.7](#) and provide their proofs. They are now mostly corollaries of the results of [Appendix A.3.5](#) and [Appendix A.3.6](#).

Theorem A.3.41. *Suppose that μ_∞^η is invariant under (A.1.15), fix a tolerance level $\varepsilon > 0$, and let $\mathcal{U}_i \equiv \mathcal{U}_i(\delta)$, $i = 1, \dots, K$, be δ -neighborhoods of the components of $\text{crit } f$. Then, for all sufficiently small $\delta, \eta > 0$, we have*

$$|\eta \log \mu_\infty^\eta(\mathcal{U}_i) + E_i - \min_j E_j| \leq \varepsilon \quad (\text{A.3.197})$$

and

$$\left| \eta \log \frac{\mu_\infty^\eta(\mathcal{U}_i)}{\mu_\infty^\eta(\mathcal{U}_j)} + E_i - E_j \right| \leq \varepsilon. \quad (\text{A.3.198})$$

More compactly, with notation as above, we have:

$$\mu_\infty^\eta(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right). \quad (\text{A.3.199})$$

Proof. Note that if μ_∞^η is invariant for (A.1.15), it is *a fortiori* invariant for the accelerated process $(x_n^\eta)_{n \geq 0}$. This result is then a direct consequence of [Proposition A.3.31](#) ([Appendix A.3.5](#)). $\blacksquare$

Theorem A.42. *Suppose that μ_∞^η is invariant under (A.1.15), and let $\mathcal{K}$ be a non-minimizing component of f . Then, with notation as in [Theorem 4.3](#), there exists a minimizing component $\mathcal{K}'$ of f and a positive constant $c \equiv c(\mathcal{K}, \mathcal{K}') > 0$ such that*

$$\frac{\mu_\infty^\eta(\mathcal{U})}{\mu_\infty^\eta(\mathcal{U}')} \leq \exp\left(-\frac{c(\mathcal{K}, \mathcal{K}') + \varepsilon}{\eta}\right) \quad (\text{A.3.200})$$

for all sufficiently small $\eta > 0$ and all sufficiently small neighborhoods $\mathcal{U}$ and $\mathcal{U}'$ of $\mathcal{K}$ and $\mathcal{K}'$ respectively. In particular, in the limit $\eta \rightarrow 0$, we have $\mu_\infty^\eta(\mathcal{U}) \rightarrow 0$.

Proof. Let $\mathcal{K} = \mathcal{K}_i$ be a non-minimizing component. By [Lemma A.3.38](#), there exists $j \in I$ such that $E_j < E_i$. The statement then follows from [Theorem 4.3](#). $\blacksquare$

Theorem A.43. *Suppose that μ_∞^η is invariant under (A.1.15), fix a tolerance level $\delta > 0$, and let $\mathcal{U} \equiv \mathcal{U}(\delta)$ be a δ -neighborhood of $\text{crit } f$. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$, we have:*

$$\mu_\infty^\eta(\mathcal{U}) \geq 1 - e^{-c/\eta}. \quad (\text{A.3.201})$$

Proof. We actually show a slightly stronger result. Define $J := \{i \in I : \mathcal{K}_i \text{ is minimizing}\}$. We prove that there exists $\mathcal{U}$ neighborhood of $\bigcup_{i \in J} \mathcal{K}_i$, a constant $c > 0$, such that, for all η small enough,

$$\mu_\infty^\eta \left(\mathbb{R}^d \setminus \bigcup_{i \in J} \mathcal{C}_i^\delta \right) \leq e^{-\frac{c}{\eta}}. \quad (\text{A.3.202})$$

This is then a consequence of [Proposition A.3.40](#) ([Appendix A.3.6](#)) with $J \leftarrow J$. Note J contain the ground states since they are minimizing by [Lemma A.3.38](#). ■

Theorem A.44. Suppose that μ_∞^η is invariant under (A.1.15), fix a tolerance level $\delta > 0$, and let $\mathcal{U}_0 \equiv \mathcal{U}_0(\delta)$ be a δ -neighborhood of the system's ground state $\mathcal{C}_0$. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$, we have:

$$\mu_\infty^\eta(\mathcal{U}_0) \geq 1 - e^{-c/\eta}. \quad (\text{A.3.203})$$

Proof. It suffices to apply [Proposition A.3.40](#) ([Appendix A.3.6](#)) with J as the set of ground states $\arg \min_{i \in I} E_i$. They are necessarily minimizing by [Lemma A.3.38](#). ■

A.3.8 Extension: the mean occupation measures

We now state and prove analogue of our main results for the mean occupation measure μ_n .

For this we will need a strengthened version of [Assumption A.3](#), viz.

Assumption A.10. The signal-to-noise ratio of G satisfies

$$\frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \rightarrow \infty \quad \text{as} \quad \|x\| \rightarrow \infty. \quad (\text{A.3.204})$$

In this section, we will posit that [Assumption A.10](#) holds in addition to [Assumptions A.1](#) and [A.2](#) and (A.3.121). We then have the following series of results for the occupation measure μ_n of (A.1.15).

Theorem A.45. Fix a tolerance level $\varepsilon > 0$, and let $\mathcal{U}_i \equiv \mathcal{U}_i(\delta)$, $i = 1, \dots, K$, be δ -neighborhoods of the components of $\text{crit } f$. Then, for all sufficiently small $\delta, \eta > 0$ and large enough n , we have

$$|\eta \log \mu_n(\mathcal{U}_i) + E_i - \min_j E_j| \leq \varepsilon \quad (\text{A.3.205})$$

and

$$\left| \eta \log \frac{\mu_n(\mathcal{U}_i)}{\mu_n(\mathcal{U}_j)} + E_i - E_j \right| \leq \varepsilon. \quad (\text{A.3.206})$$

More compactly, with notation as above, we have:

$$\mu_n(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right). \quad (\text{A.3.207})$$

Theorem A.46. Let $\mathcal{K}$ be a non-minimizing component of f . Then, with notation as in [Theorem 4.3](#), there exists a minimizing component $\mathcal{K}'$ of f and a positive constant $c \equiv c(\mathcal{K}, \mathcal{K}') > 0$ such that

$$\frac{\mu_n(\mathcal{U})}{\mu_n(\mathcal{U}')} \leq \exp\left(-\frac{c(\mathcal{K}, \mathcal{K}') + \varepsilon}{\eta}\right) \quad (\text{A.3.208})$$

for all sufficiently small $\eta > 0$, n large enough and all sufficiently small neighborhoods $\mathcal{U}$ and $\mathcal{U}'$ of $\mathcal{K}$ and $\mathcal{K}'$ respectively.

Theorem A.47. Fix a tolerance level $\delta > 0$, and let $\mathcal{U} \equiv \mathcal{U}(\delta)$ be a δ -neighborhood of $\text{crit } f$. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$ and large enough n , we have:

$$\mu_n(\mathcal{U}) \geq 1 - e^{-c/\eta}. \quad (\text{A.3.209})$$

Theorem A.48. Fix a tolerance level $\delta > 0$, and let $\mathcal{U}_0 \equiv \mathcal{U}_0(\delta)$ be a δ -neighborhood of the system's ground state $\mathcal{C}_0$. Then there exists a constant $c \equiv c_\delta > 0$ such that, for all sufficiently small $\eta > 0$ and large enough n , we have:

$$\mu_n(\mathcal{U}_0) \geq 1 - e^{-c/\eta}. \quad (\text{A.3.210})$$

We begin with a preliminary lemma which shows that, under [Assumption A.10](#), the sequence of mean occupation measure $(\mu_n)_{n \geq 0}$ is tight (see e.g., [Kallenberg \(2021, Chap. 23\)](#)).

Lemma A.3.49. The sequence of mean occupation measures $(\mu_n)_{n \geq 0}$ is tight.

The proof of this lemma first follows the proof of [Lemma A.3.19](#) and then relies on the same reasoning as the proof of [Douc et al. \(2018, Thm. 12.3.3\)](#).

Proof. By [Lemma A.3.17](#), there exists $\mathcal{K} \subset \mathbb{R}^d$ compact, $\eta_0 > 0$, $c > 0$ such that, for any $\eta \leq \eta_0$, $x_0 = x \notin \mathcal{K}$,

$$V_\infty(x_1) - V_\infty(x) \leq \eta \left(\frac{\|U(x, \omega_0)\|^2}{\sigma_\infty^2(f(x))} - \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \right). \quad (\text{A.3.211})$$

Passing to the expectation yields that, for any $x \notin \mathcal{K}$,

$$\mathbb{E}_x[V_\infty(x_1)] - V_\infty(x) \leq \eta \left(\frac{\mathbb{E}_x[\|U(x, \omega_0)\|^2]}{\sigma_\infty^2(f(x))} - \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} \right). \quad (\text{A.3.212})$$

Applying [Corollary A.3.16](#) with $X \leftarrow \frac{U(x, \omega_0)}{\sqrt{\sigma_\infty^2(f(x))}}$ (the conditions of application are verified from [Assumption A.2\(c\)](#)) yields that

$$\mathbb{E}_x[V_\infty(x_1)] - V_\infty(x) \leq -\eta \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} + \eta \times 16d \log 6. \quad (\text{A.3.213})$$

Hence, for any $x \in \mathbb{R}^d$, we have

$$\begin{aligned} \mathbb{E}_x[V_\infty(x_1)] - V_\infty(x) &\leq \mathbb{1}\{x \in \mathcal{K}\} \left(\sup_{x' \in \mathcal{K}} \mathbb{E}_{x'}[V_\infty(x')] - \inf_{\mathbb{R}^d} V_\infty \right) \\ &\quad - \mathbb{1}\{x \notin \mathcal{K}\} \left(\eta \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} - \eta \times 16d \log 6 \right), \end{aligned} \quad (\text{A.3.214})$$

or, after rearranging,

$$\begin{aligned} \mathbb{E}_x[V_\infty(x_1)] - \mathbb{1}\{x \in \mathcal{K}\} \left(\sup_{x' \in \mathcal{K}} \mathbb{E}_{x'}[V_\infty(x')] - \inf_{\mathbb{R}^d} V_\infty \right) \\ + \mathbb{1}\{x \notin \mathcal{K}\} \left(\eta \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} - \eta \times 16d \log 6 \right) \leq V_\infty(x). \end{aligned} \quad (\text{A.3.215})$$

Since the function

$$\begin{aligned} x \mapsto & -\mathbb{1}\{x \in \mathcal{K}\} \left(\sup_{x' \in \mathcal{K}} \mathbb{E}_{x'}[V_\infty(x')] - \inf_{\mathbb{R}^d} V_\infty \right) \\ & + \mathbb{1}\{x \notin \mathcal{K}\} \left(\eta \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2(f(x))} - \eta \times 16d \log 6 \right) \end{aligned} \quad (\text{A.3.216})$$

is measurable, lower-bounded and goes to infinity as $\|x\| \rightarrow \infty$ by [Assumption A.10](#), one can then apply the same computations as in the proof of [Douc et al. \(2018, Thm. 12.3.3\)](#) to obtain that the sequence of occupation measures $(\mu_n)_{n \geq 0}$ is tight. ■

We now prove [Theorem A.45](#) by adapting the proof of [Theorem A.3.41](#). Since the process is exactly the same for [Theorem A.46](#), [Theorem A.47](#), and [Theorem A.48](#), we omit their proofs.

Proof of Theorem A.45. We show that, for sufficiently small $\delta, \eta > 0$, for any $i \in I$, we have that

$$\begin{aligned} \exp\left(-\frac{E_i - \min_j E_j + \varepsilon}{\eta}\right) & \leq \liminf_{n \rightarrow \infty} \mu_n(\mathcal{U}_i) \\ & \leq \limsup_{n \rightarrow \infty} \mu_n(\mathcal{U}_i) \\ & \leq \exp\left(-\frac{E_i - \min_j E_j - \varepsilon}{\eta}\right) \end{aligned} \quad (\text{A.3.217})$$

and the results in the statement will follow with 2ε in place of ε .

We apply [Proposition A.3.31](#) ([Appendix A.3.5](#)): take δ small enough so [Proposition A.3.31](#) can be applied with both the neighborhoods $\mathcal{U}_1, \dots, \mathcal{U}_K$ and $\text{cl}\mathcal{U}_1, \dots, \text{cl}\mathcal{U}_K$. One then obtain $\eta_0 > 0$ such that, for all $0 < \eta < \eta_0$ and any μ_∞^η invariant probability measure for $(x_n^\eta)_n$, for any $i \in I$,

$$\begin{aligned} \mu_\infty^\eta(\mathcal{U}_i) & \geq \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j \in I} E(\mathcal{K}_j)}{\eta} - \frac{\varepsilon}{\eta}\right) \\ \mu_\infty^\eta(\text{cl}\mathcal{U}_i) & \leq \exp\left(-\frac{E(\mathcal{K}_i) - \min_{j \in I} E(\mathcal{K}_j)}{\eta} + \frac{\varepsilon}{\eta}\right). \end{aligned} \quad (\text{A.3.218})$$

We now prove that [\(A.3.217\)](#) holds. Fix $i \in I$. By [Lemma A.3.49](#), the sequence of mean occupation measures $(\mu_n)_{n \geq 0}$ is tight, so that, by Prohorov theorem ([Kallenberg, 2021, Thm. 23.2](#)), it is sequentially compact for the weak topology, or, in other terms, for the convergence in distribution. Therefore, $(\mu_n)_{n \geq 0}$ admits a weak accumulation point which is a probability distribution and that we denote by ν . Applying Portmanteau theorem ([Kallenberg, 2021, Thm. 5.25](#)) to the open set $\mathcal{U}_i$ and the closed set $\text{cl}\mathcal{U}_i$ yields that

$$\nu(\mathcal{U}_i) \leq \liminf_{n \rightarrow \infty} \mu_n(\mathcal{U}_i) \leq \limsup_{n \rightarrow \infty} \mu_n(\text{cl}\mathcal{U}_i) \leq \nu(\text{cl}\mathcal{U}_i). \quad (\text{A.3.219})$$

Since $(x_n)_{n \geq 0}$, the sequence of iterates of SGD, is (weak) Feller by [Lemma A.3.18](#), ν is actually invariant for $(x_n)_{n \geq 0}$, by, e.g., [Douc et al. \(2018, Prop. 12.3.1\)](#), and *a fortiori* invariant for the accelerated process $(x_n^\eta)_{n \geq 0}$. Combining [\(A.3.218\)](#) with [\(A.3.219\)](#) gives the result [\(A.3.217\)](#). ■

A.4 POTENTIAL FOR THE INVARIANT MEASURE

A.4.1 Gaussian noise

Though it does not formally fit into our setting, let us first begin with the case where the noise is Gaussian. Since it is unbounded, our assumptions are not satisfied and our theorems describing the invariant measure do not apply. However, all the objects we consider are still well-defined, and, in that case, it is possible to compute the E_i explicitly. Moreover, this section serves as a blueprint for the truncated Gaussian case of the next section.

Assume that, for every $x \in \mathbb{R}^d$, $U(x, \omega)$ follows a centered Gaussian distribution with covariance $\sigma^2(f(x))I_d$ for some continuous function $\sigma^2 : \mathbb{R} \rightarrow (0, +\infty)$.

Akin to V_∞ in [Appendix A.3](#), a key role is played by the function $V : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$V(x) := 2\alpha(f(x)) \quad \text{with} \quad \alpha : \mathbb{R} \rightarrow \mathbb{R} \quad \text{a primitive of} \quad 1/\sigma^2. \quad (\text{A.4.1})$$

Since the noise is Gaussian, the Lagrangian and Hamiltonian have explicit expressions: for every $x, p, v \in \mathbb{R}^d$,

$$\mathcal{H}(x, p) = -\langle \nabla f(x), p \rangle + \frac{1}{2} \sigma^2(f(x)) \|p\|^2 \quad (\text{A.4.2a})$$

$$\mathcal{L}(x, v) = \frac{\|v + \nabla f(x)\|^2}{2\sigma^2(f(x))}. \quad (\text{A.4.2b})$$

This expression of $\mathcal{L}$ make it clear that the action function penalizes the deviation of a path from the flow: for a path $\gamma \in \mathcal{C}([0, T])$ for some $T > 0$,

$$\mathcal{S}_T(\gamma) = \int_0^T \frac{\|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))} dt. \quad (\text{A.4.3})$$

The computation of the E_i relies on the following observation. Take a path $\gamma \in \mathcal{C}([0, T])$ for some $T > 0$ and consider φ defined by $\varphi_t = \gamma_{T-t}$ for $t \in [0, T]$. Then, the action cost of φ is given by,

$$\begin{aligned} \mathcal{S}_T(\varphi) &= \int_0^T \frac{\|-\dot{\gamma}_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))} dt \\ &= \int_0^T \frac{\|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2}{2\sigma^2(f(\gamma_t))} dt - \int_0^T \frac{2\langle \dot{\gamma}_t, \nabla f(\gamma_t) \rangle}{\sigma^2(f(\gamma_t))} dt \\ &= \mathcal{S}_T(\gamma) - \int_0^T \langle \dot{\gamma}_t, \nabla V(\gamma_t) \rangle dt, \end{aligned} \quad (\text{A.4.4})$$

since $\nabla V(x) = 2\alpha'(f(x))\nabla f(x)$. Therefore, we get that

$$\mathcal{S}_T(\varphi) = \mathcal{S}_T(\gamma) - (V(\gamma_T) - V(\gamma_0)). \quad (\text{A.4.5})$$

Take $i, j \in I$. This equality then translates to a relation between $\mathcal{B}_{i,j}$ and $\mathcal{B}_{j,i}$: considering $\gamma \in \mathcal{C}([0, T])$ such that $\gamma_0 \in \mathcal{K}_i$, $\gamma_T \in \mathcal{K}_j$ and taking the infimum over all such paths, we get that

$$\mathcal{B}_{j,i} \leq \mathcal{B}_{i,j} + (V_j - V_i). \quad (\text{A.4.6})$$

where, since f is constant on $\mathcal{K}_i$ and $\mathcal{K}_j$, we denote by V_i and V_j the values of V on $\mathcal{K}_i$ and $\mathcal{K}_j$ respectively.

Reversing the roles of i and j and applying the same argument shows that this inequality is an equality:

$$\mathcal{B}_{j,i} + V_j = \mathcal{B}_{i,j} + V_i. \quad (\text{A.4.7})$$

Denote by $C_{i,j}$ this common value. Crucially, $C_{i,j}$ is symmetric in i and j .

Consider now t^* a minimum weight spanning tree on the complete but now undirected graph on I with weights $(C_{i,j})_{i,j}$. We show that the minima in the E_i are attained t^* , or more precisely, a directed version of it.

Fix i and consider $t \in \mathcal{T}_i$ a spanning tree rooted at i . We have that, since any node j is the origin of exactly one edge in t ,

$$\begin{aligned} \sum_{(j \rightarrow l) \in t} \mathcal{B}_{j,l} + \sum_{j \in I} V_j &= \sum_{(j \rightarrow l) \in t} (\mathcal{B}_{j,l} + V_j) + V_i \\ &= \sum_{(j \rightarrow l) \in t} C_{j,l} + V_i. \end{aligned} \quad (\text{A.4.8})$$

But by definition of t^* , this sum is at least greater than $\sum_{(j \leftrightarrow l) \in t^*} C_{j,l}$ so that we have

$$\sum_{(j \rightarrow l) \in t} \mathcal{B}_{j,l} + \sum_{j \in I} V_j \geq \sum_{(j \leftrightarrow l) \in t^*} C_{j,l} + V_i, \quad (\text{A.4.9})$$

and taking t_i an oriented version of t^* rooted at i , the equality is attained. Therefore, we have that

$$E_i = \sum_{(j \rightarrow l) \in t_i} \mathcal{B}_{j,l} = \sum_{(j \leftrightarrow l) \in t^*} C_{j,l} - \sum_{j \in I} V_j + V_i, \quad (\text{A.4.10})$$

or, in short, $E_i = V_i + c$ where $c > 0$ is independent of i .

Therefore, the mass distribution over critical points is governed by a Gibbs measure with potential V .

Let us now mention two particular cases.

- If σ^2 is constant, then $V = \frac{2f}{\sigma^2}$.
- If σ^2 is linear, i.e., of the form $\sigma^2(f(x)) = \sigma_1^2(f(x) + \sigma_0^2)$, then $V = \frac{2}{\sigma_1^2} \log(f + \sigma_0^2)$.

A.4.2 Truncated Gaussian noise

To fit into our theoretical framework, we consider truncated Gaussian noise instead. The general outline of the proof but with added steps to handle the error due to the truncation. In particular, one must show that, without loss of generality, we can only consider paths whose derivative has the same norm as the gradient of f . This is done with [Freidlin and Wentzell \(1998b, Chap. 4, Lem. 3.1\)](#) that we adapt to our setting.

Assume that $U(x, \omega)$ follows a centered Gaussian distribution with covariance $\sigma^2(f(x))I_d$ conditioned on being in $\mathbb{B}(0, R(x))$ for some $R(x) > 0$.

As in [Definition A.3](#), we define $V(x) = 2\alpha(f(x))$ with $\alpha' = \frac{1}{\sigma^2}$ and denote by V_i the value taken by V on $\mathcal{K}_i$.

Consider some $0 < \delta \leq \frac{1}{2}$ and assume that

$$\sup_{x \in \mathbb{R}^d} 2^{d+4}(d+1)e^{-\frac{R^2}{16\sigma^2}} \leq \delta, \quad (\text{A.4.11})$$

so that the error term $2E(\sigma^2(f(x)), R(x))$ in [Lemma A.5.2](#) is bounded by δ .

Moreover, assume that, for any $x \in \mathbb{R}^d$

$$\|\nabla f(x)\| \leq \frac{R(x)}{8} \quad (\text{A.4.12})$$

Lemma A.4.1. *Consider $\gamma \in \mathcal{C}([0, T])$. Then, there exists $\tilde{\gamma} \in \mathcal{C}([0, S])$ a reparametrization of γ such that, for any $t \in [0, S]$,*

$$\|\dot{\tilde{\gamma}}_s\| = \|\nabla f(\tilde{\gamma}_s)\|. \quad (\text{A.4.13})$$

and

$$\mathcal{S}_T(\gamma) \geq \int_0^S \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2(1 + \delta)\sigma^2(f(\tilde{\gamma}_s))} ds. \quad (\text{A.4.14})$$

Proof. By the proof [Freidlin and Wentzell \(1998b, Chap. 4, Lem. 3.1\)](#), there exists $t(s)$ change of time such that, with $\tilde{\gamma}_s = \gamma_{t(s)}$, $\|\dot{\tilde{\gamma}}_s\| = \|\nabla f(\tilde{\gamma}_s)\|$.

We have that

$$\mathcal{S}_T(\gamma) = \int_0^{t^{-1}(T)} \dot{t}(s) \mathcal{L}(\tilde{\gamma}_s, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s) ds, \quad (\text{A.4.15})$$

so it suffices to bound $\mathcal{L}(\tilde{\gamma}_s, \dot{\tilde{\gamma}}_s)$ from below: by definition, we have

$$\begin{aligned} \mathcal{L}(\tilde{\gamma}_s, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s) &\geq \sup \left\{ \langle p, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s) \rangle - \bar{\mathcal{H}}(\tilde{\gamma}_s, p) : \|p\| \leq \frac{R(\tilde{\gamma}_s)}{2\sigma^2(f(\tilde{\gamma}_s))} \right\} \\ &\geq \sup \left\{ \langle p, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s) \rangle - (1 + \delta) \frac{\sigma^2(f(\tilde{\gamma}_s))}{2} \|p\|^2 : \|p\| \leq \frac{R(\tilde{\gamma}_s)}{2\sigma^2(f(\tilde{\gamma}_s))} \right\}, \end{aligned} \quad (\text{A.4.16})$$

by [Lemma A.5.2](#). Applying [Lemma A.5.3](#) with $v \leftarrow (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s$, $w \leftarrow \nabla f(\tilde{\gamma}_s)$ and $\lambda \leftarrow (\dot{t}(s))^{-1}$, now exactly yields, for almost all s ,

$$\begin{aligned} \mathcal{L}(\tilde{\gamma}_s, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s) &\geq (\dot{t}(s))^{-1} \sup \left\{ \langle p, \dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s) \rangle - (1 + \delta) \frac{\sigma^2(f(\tilde{\gamma}_s))}{2} \|p\|^2 : \|p\| \leq \frac{R(\tilde{\gamma}_s)}{2\sigma^2(f(\tilde{\gamma}_s))} \right\} \\ &= (\dot{t}(s))^{-1} \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2(1 + \delta)\sigma^2(f(\tilde{\gamma}_s))}, \end{aligned} \quad (\text{A.4.17})$$

since $\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\| \leq \frac{R(\tilde{\gamma}_s)}{2}$. ■

Lemma A.4.2. *With this setting, for any $i \in I$*

$$E_i = \sum_{(j \leftrightarrow l) \in I^*} C_{j,l} - \sum_{j \in I} V_j + V_i, \quad (\text{A.4.18})$$

Proof. Consider $\gamma \in \mathcal{C}([0, T])$ such that $\gamma_0 \in \mathcal{K}_i$, $\gamma_T \in \mathcal{K}_j$ and $\mathcal{S}_T(\gamma) < +\infty$. Then, by the previous lemma [Lemma A.4.1](#), there exists $\tilde{\gamma} \in \mathcal{C}([0, S])$ a reparametrization of γ such that, for any $t \in [0, S]$,

$$\begin{aligned} \mathcal{S}_T(\gamma) &\geq \int_0^S \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2(1 + \delta)\sigma^2(f(\tilde{\gamma}_s))} ds \\ &= \int_0^S \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2(1 + \delta)\sigma^2(f(\tilde{\gamma}_s))} ds + \frac{V(\gamma_T) - V(\gamma_0)}{1 + \delta}, \end{aligned} \quad (\text{A.4.19})$$

where we performed the same computations as above in [Appendix A.4.1](#). Considering the path $(\tilde{\gamma}_{s-s})_{s \in [0, S]}$ and invoking the upper-bound on the Lagrangian from [Lemma A.5.2](#) with $-\tilde{\gamma}_s + \nabla f(\tilde{\gamma}_s)$ which still has norm less than $R(\tilde{\gamma}_s)/4$, we get that

$$\mathcal{S}_T(\gamma) \geq \frac{1-\delta}{1+\delta} \mathcal{B}_{j,i} + \frac{V(\gamma_T) - V(\gamma_0)}{1+\delta}. \quad (\text{A.4.20})$$

The result now follows from the same computations as in [Appendix A.4.1](#). ■

A.4.3 Local dependencies

Under local assumptions similar to [Mori et al. \(2022\)](#), we demonstrate how the modelling of the noise influences the invariant measure.

Lemma A.4.3. *Consider $\mathcal{K}_i$ minimizing, $\sigma^2: \mathbb{R} \rightarrow (0, +\infty)$ continuous, and take $\alpha: \mathbb{R} \rightarrow \mathbb{R}$ such that $\alpha' = \frac{1}{\sigma^2}$. Assume that H^* is a positive definite matrix such that, locally near $\mathcal{K}_i$, it holds that:*

$$\alpha(f(x)) = \sum_{\lambda \in \text{eig } H^*} g^\lambda(x_\lambda), \quad (\text{A.4.21})$$

where x_λ denotes the orthogonal projection of x on the eigenspace of the eigenvalue λ and where $g^\lambda: \mathbb{R}^d \rightarrow \mathbb{R}$ is continuously differentiable. Define the potential $V: \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$V(x) = \sum_{\lambda \in \text{eig } H^*} \frac{2g^\lambda(x_\lambda)}{\lambda}. \quad (\text{A.4.22})$$

If we have the anisotropic subGaussian bound: for x in a neighborhood of $\mathcal{K}_i$, $p \in \overline{\mathbb{B}}(0, \|V(x)\|)$,

$$\tilde{\mathcal{H}}(x, p) \leq \frac{\sigma^2(f(x))}{2} \langle p, H^* p \rangle, \quad (\text{A.4.23})$$

then there is $\delta > 0$ such that, for all $j \neq i$, any $0 < \delta' \leq \delta$,

$$E_j \geq \min\{V(x) - V_i : x, d(x, \mathcal{K}_i) = \delta'\} > 0, \quad (\text{A.4.24})$$

where V_i is the value of V on $\mathcal{K}_i$.

Moreover, there exists $R > \delta$ such that, if there exists $c > 0$, $s^2 > 0$, such that, for all $x \in \overline{\mathbb{B}}(0, R) \setminus \mathcal{U}_\delta(\mathcal{K}_i)$, $v \in \mathbb{B}(\nabla f(x), c)$,

$$\tilde{\mathcal{L}}(x, v) \leq \frac{\|v\|}{2s^2}, \quad (\text{A.4.25})$$

then there exists $C > 0$ that depends only on r, R, c, f restricted to $\mathbb{R}^d \setminus \mathcal{U}_r(\mathcal{K}_i)$ such that

$$E_i \leq \frac{C}{s^2}. \quad (\text{A.4.26})$$

The assumptions on the Hamiltonian and the Lagrangian roughly say that the noise share some similarities with Gaussian distributions with the prescribed variances. In particular, they are satisfied in the (truncated) Gaussian case, as shown in [Lemma A.5.2](#).

Moreover, a takeaway of this lemma is that, if σ^2 or the eigenvalues of H^* are small enough, then $\mathcal{K}_i$ must be the ground state even if it may not be the global minimum of V . This lemma is general enough to handle non-constant variance: a notable example is when $\sigma^2(f(x))$ is linear in $f(x)$ and where the resulting potential V then depends logarithmically on the value f .

Proof. Denote $P_\lambda \in \mathbb{R}^{d \times d}$ the orthogonal projection on the eigenspace of H^* associated with the eigenvalue λ . V can thus be rewritten as

$$V(x) = \sum_{\lambda \in \text{eig } H^*} \frac{2g^\lambda(P_\lambda x)}{\lambda}, \quad (\text{A.4.27})$$

so that its gradient is given by

$$\nabla V(x) = \sum_{\lambda \in \text{eig } H^*} \frac{2P_\lambda \nabla g^\lambda(P_\lambda x)}{\lambda}. \quad (\text{A.4.28})$$

In particular, we obtain that for x close enough to $\mathcal{K}_i$, using the orthogonality of the projections,

$$\begin{aligned} \bar{\mathcal{H}}(x, \nabla V(x)) &\leq \frac{\sigma^2(f(x))}{2} \langle \nabla V(x), H^* \nabla V(x) \rangle \\ &= \frac{\sigma^2(f(x))}{2} \sum_{\lambda \in \text{eig } H^*} \frac{4\|P_\lambda \nabla g^\lambda(P_\lambda x)\|^2}{\lambda} \\ &= \frac{\sigma^2(f(x))}{2} \langle \nabla V(x), 2\nabla(\alpha \circ f)(x) \rangle \\ &= \langle \nabla V(x), \nabla f(x) \rangle. \end{aligned} \quad (\text{A.4.29})$$

Therefore, for x close enough to $\mathcal{K}_i$, we have that

$$\mathcal{H}(x, \nabla V(x)) = -\langle \nabla V(x), \nabla f(x) \rangle + \bar{\mathcal{H}}(x, \nabla V(x)) \leq 0. \quad (\text{A.4.30})$$

For x close to $\mathcal{K}_i$, let us compute $\langle \nabla V(x), \nabla f(x) \rangle$:

$$\begin{aligned} \langle \nabla V(x), \nabla f(x) \rangle &= \sigma^2(f(x)) \langle \nabla V(x), \nabla(\alpha \circ f)(x) \rangle \\ &= \sigma^2(f(x)) \sum_{\lambda \in \text{eig } H^*} \frac{2\|P_\lambda \nabla g^\lambda(P_\lambda x)\|}{\lambda}, \end{aligned} \quad (\text{A.4.31})$$

which is (strictly) positive in a small neighborhood of $\mathcal{K}_i$ (excluding $\mathcal{K}_i$ itself). Therefore, V is decreasing on trajectories of the flow in this neighborhood. With the same proof as in [Lemma A.3.35](#), we deduce that there exists $\delta > 0$ such that $\arg \min_{x \in \mathcal{U}_\delta(\mathcal{K}_i)} V(x) = \mathcal{K}_i$. Moreover, take $\delta > 0$ small enough such that [\(A.4.30\)](#) holds on $\mathcal{U}_\delta(\mathcal{K}_i)$, $\mathcal{U}_\delta(\mathcal{K}_i) \cap \text{crit } f = \mathcal{K}_i$ and the trajectories of the flow started in $\mathcal{U}_\delta(\mathcal{K}_i)$ stay converge to $\mathcal{K}_i$. We now proceed as in the proof of [Lemma A.3.39](#). Take $\delta' \leq \delta/2$, $\mathcal{U} := \mathcal{U}'_\delta(\mathcal{K}_i)$ and $\Delta := \min\{V(x) - V_i : x \in \mathbb{R}^d, d(x, \mathcal{K}_i) = \delta'\}$, which is positive by definition. Fix $j \neq i$: we show that $B_{i,j} \geq \Delta$. Consider some $T > 0$ and $\gamma \in \mathcal{C}([0, T], \mathbb{R}^d)$ such that $\gamma_0 \in \mathcal{K}_i$ and $\gamma_T \in \mathcal{K}_j$. By definition of δ and by continuity of γ and $d(\cdot, \mathcal{K}_i)$, there exists $t \in [0, T]$ such that $d(\gamma_t, \mathcal{K}_i) = \delta'$. Therefore, we have that

$$\mathcal{S}_{0,T}(\gamma) \geq \mathcal{S}_{0,t}(\gamma) = \int_0^t \mathcal{L}(\gamma_s, \dot{\gamma}_s) ds, \quad (\text{A.4.32})$$

and therefore, by definition of $\mathcal{L}$ as the conjugate of $\mathcal{H}$, we obtain that

$$\mathcal{S}_{0,T}(\gamma) \geq \int_0^t \langle \dot{\gamma}_s, \nabla V(\gamma_s) \rangle - \bar{\mathcal{H}}(\gamma_s, \nabla V(\gamma_s)) ds \geq \int_0^t \langle \dot{\gamma}_s, \nabla V(\gamma_s) \rangle, \quad (\text{A.4.33})$$

where we used (A.4.30) in the last inequality. Thus, we get

$$\mathcal{S}_{0,T}(\gamma) \geq V(\gamma_T) - V(\gamma_0) \geq \Delta, \quad (\text{A.4.34})$$

with any $\delta' \leq \delta/2$ (and therefore δ in the statement corresponds to $\delta/2$ here).

Finally, it remains to show that, any $l \neq i$, $E_l \geq \Delta$. Consider any tree rooted at l : it must have an edge of the form $i \rightarrow j$ for some $j \in I$ and therefore the sum of its weights will be at least $\mathcal{B}_{i,j} \geq \Delta$. Hence, since it holds for any such tree, we have that $E_l \geq \Delta$.

We now prove the second part of the lemma. δ was chosen small enough so that it is possible to find $R > \delta$ such that $\mathbb{B}(0, R) \setminus \mathcal{U}_\delta(\mathcal{K}_i)$ contains all the $\mathcal{K}_j$ for $j \neq i$ and not $\mathcal{K}_i$. The assumption on the Lagrangian implies that, for any $x \in \mathbb{B}(0, R) \setminus \mathcal{U}_\delta(\mathcal{K}_i)$, $v \in \mathbb{B}(0, c)$,

$$\mathcal{L}(x, v) \leq \frac{\|v + \nabla f(x)\|^2}{2s^2}. \quad (\text{A.4.35})$$

Fix $j \neq i$ and take $\gamma \in \mathcal{C}^1([0, T], \mathbb{R}^d)$ such that $\gamma_0 \in \mathcal{K}_j$, $x := \gamma_T \in \mathcal{U}_\delta(\mathcal{K}_i)$ and γ remains in $\overline{\mathbb{B}}(0, R) \setminus \mathcal{U}_\delta(\mathcal{K}_i)$. Without loss of generality, at the expense of replacing γ by a reparametrization, we can assume that $\dot{\gamma}_t$ in $\mathbb{B}(0, c)$ for all $t \in [0, T]$. Then, we have that

$$\mathcal{S}_{0,T}(\gamma) \leq \int_0^T \mathcal{L}(\gamma_s, \dot{\gamma}_s) ds \leq \int_0^T \frac{\|\dot{\gamma}_s + \nabla f(\gamma_s)\|^2}{2s^2} ds, \quad (\text{A.4.36})$$

which depends only on the value of ∇f outside of $\mathcal{U}_{\delta/2}(\mathcal{K}_i)$. But, by the same reasoning as in Lemma A.3.37, since the flow started at x converges to $\mathcal{K}_i$, we have that $\mathcal{B}(\{x\}, \mathcal{K}_i) = 0$. Therefore, we have that,

$$\mathcal{B}_{j,i} \leq \int_0^T \frac{\|\dot{\gamma}_s + \nabla f(\gamma_s)\|^2}{2s^2} ds, \quad (\text{A.4.37})$$

and, taking the maximum of such quantities over all $j \neq i$, we obtain $C > 0$ such, for any $j \neq i$,

$$\mathcal{B}_{j,i} \leq \frac{C}{s^2}. \quad (\text{A.4.38})$$

To conclude on the value of E_i , we consider the tree rooted at i made of all the edges (j, i) for $j \neq i$. It has weight at most $(K-1)C/s^2$ and therefore, we have that $E_i \leq (K-1)C/s^2$. ■

A.5 AUXILIARY RESULTS

A.5.1 Truncated Gaussian distribution

Lemma A.5.1. Consider $X \sim \mathcal{N}(0, \Sigma)$ a multivariate Gaussian distribution $\Sigma \in \mathbb{R}^{d \times d}$ positive definite. For $R > 0$, define the truncated Gaussian random variable (r.v.) X_R by conditioning X to the ball $\overline{\mathbb{B}}(0, R)$. Define

$$\bar{\mathcal{H}}(p) := \log \mathbb{E} \left[e^{\langle p, X_R \rangle} \right] \quad (\text{A.5.1})$$

and

$$\tilde{E}(\Sigma, \tilde{R}) := e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} (\text{tr } \Sigma + \|\Sigma\|) 2^{d+3} \quad (\text{A.5.2})$$

Then, for $\tilde{R} > 0$ such that

$$\tilde{R} \geq \sqrt{\|\Sigma\|(2d+4)\log 2}, \quad (\text{A.5.3})$$

it holds that, for any $p \in \mathbb{R}^d$ such that $\|\Sigma p\| \leq R - \tilde{R}$,

$$|\tilde{\mathcal{H}}(p) - \frac{1}{2}\langle p, \Sigma p \rangle| \leq \frac{1}{2}\tilde{E}(\Sigma, \tilde{R})\|p\|^2 \quad (\text{A.5.4a})$$

$$\|\nabla \tilde{\mathcal{H}}(p) - \Sigma p\| \leq \tilde{E}(\Sigma, \tilde{R})\|p\| \quad (\text{A.5.4b})$$

$$\|\text{Hess } \tilde{\mathcal{H}}(p) - \Sigma\| \leq \tilde{E}(\Sigma, \tilde{R}). \quad (\text{A.5.4c})$$

Proof. Define, for A a measurable set,

$$Z(A) = \int_A e^{-\frac{1}{2}x \cdot \Sigma^{-1}x} dx, \quad (\text{A.5.5})$$

and, for convenience, $\mathcal{K} := \overline{\mathbb{B}}(0, R)$. For notational convenience, in this proof, we will denote the inner product between two vectors x and p with a simple dot $x \cdot p$. We have that

$$\begin{aligned} \mathbb{E}[e^{X_R \cdot p}] &= \frac{e^{\frac{1}{2}p \cdot \Sigma p}}{Z(\mathcal{K})} \int_{\mathcal{K}} e^{-\frac{1}{2}(x - \Sigma p) \cdot \Sigma^{-1}(x - \Sigma p)} dx \\ &= \frac{e^{\frac{1}{2}p \cdot \Sigma p}}{Z(\mathcal{K})} \int_{\mathcal{K} - \Sigma p} e^{-\frac{1}{2}x' \cdot \Sigma^{-1}x'} dx' \\ &= e^{\frac{1}{2}p \cdot \Sigma p} \frac{Z(\mathcal{K} - \Sigma p)}{Z(\mathcal{K})}, \end{aligned} \quad (\text{A.5.6})$$

where we performed the change of variable $x' \leftarrow x - \Sigma p$.

Define

$$f(p) := Z(\mathcal{K} - \Sigma p) = \int_{\mathcal{K}} e^{-\frac{1}{2}(x - \Sigma p) \cdot \Sigma^{-1}(x - \Sigma p)} dx. \quad (\text{A.5.7})$$

so that

$$\tilde{\mathcal{H}}(p) = \log \mathbb{E}[e^{X_R \cdot p}] = \frac{1}{2}p \cdot \Sigma p + \log \frac{f(p)}{f(0)}. \quad (\text{A.5.8})$$

Therefore it suffices to bound $\log \frac{f(p)}{f(0)}$ and its derivatives.

Differentiating yields

$$\nabla f(p) = \int_{\mathcal{K}} (x - \Sigma p) e^{-\frac{1}{2}(x - \Sigma p) \cdot \Sigma^{-1}(x - \Sigma p)} dx \quad (\text{A.5.9a})$$

$$\text{Hess } f(p) = \int_{\mathcal{K}} (x - \Sigma p)(x - \Sigma p)^\top e^{-\frac{1}{2}(x - \Sigma p) \cdot \Sigma^{-1}(x - \Sigma p)} dx - \Sigma f(p). \quad (\text{A.5.9b})$$

Note that, by symmetry of $\overline{\mathbb{B}}(0, R)$, $\nabla f(0) = 0$.

We first bound $\text{Hess } f(p)$. Performing the change of variable $x' \leftarrow x - \Sigma p$ again gives us

$$\begin{aligned} \text{Hess } f(p) &= \int_{\mathcal{K} - \Sigma p} x' x'^\top e^{-\frac{1}{2}x' \cdot \Sigma^{-1}x'} dx' - \Sigma f(p) \\ &= \int_{\mathcal{K} - \Sigma p} (x' x'^\top - \Sigma) e^{-\frac{1}{2}x' \cdot \Sigma^{-1}x'} dx' \end{aligned}$$

$$= - \int_{\mathbb{R}^d \setminus (\mathcal{K} - \Sigma p)} (x' x'^\top - \Sigma) e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx', \quad (\text{A.5.10})$$

where we used that $\int_{\mathbb{R}^d} (x' x'^\top - \Sigma) e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' = 0$.

By definition of $\tilde{R}$, $\mathcal{K} - \Sigma p$ contains $\overline{\mathbb{B}}(0, \tilde{R})$. We now bound the operator norm of $\text{Hess } f(p)$:

$$\begin{aligned} \|\text{Hess } f(p)\| &\leq \int_{\mathbb{R}^d \setminus (\mathcal{K} - \Sigma p)} (\|x'\|^2 + \|\Sigma\|) e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' \\ &\leq \int_{\mathbb{R}^d \setminus \overline{\mathbb{B}}(0, \tilde{R})} (\|x'\|^2 + \|\Sigma\|) e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' \\ &\leq e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} \int_{\mathbb{R}^d} (\|x'\|^2 + \|\Sigma\|) e^{-\frac{1}{4} x' \cdot \Sigma^{-1} x'} dx' \\ &= e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} (\text{tr } \Sigma + \|\Sigma\|) (4\pi)^{d/2} \det(\Sigma)^{1/2}, \end{aligned} \quad (\text{A.5.11})$$

We now bound $\nabla f(p)$. The change of variable $x' \leftarrow x - \Sigma p$ yields

$$\nabla f(p) = \int_{\mathcal{K} - \Sigma p} x' e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' = \int_{\mathbb{R}^d \setminus (\mathcal{K} - \Sigma p)} x' e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx', \quad (\text{A.5.12})$$

where we used that $\int_{\mathbb{R}^d} x' e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' = 0$.

Therefore, similar computations as above yield that

$$\begin{aligned} \|\nabla f(p)\| &\leq \int_{\mathbb{R}^d \setminus (\mathcal{K} - \Sigma p)} \|x'\| e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' \\ &\leq \int_{\mathbb{R}^d \setminus \overline{\mathbb{B}}(0, \tilde{R})} \|x'\| e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' \\ &\leq e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} \int_{\mathbb{R}^d} \|x'\| e^{-\frac{1}{4} x' \cdot \Sigma^{-1} x'} dx' \\ &\leq e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} \sqrt{\text{tr } \Sigma} (4\pi)^{d/2} \det(\Sigma)^{1/2}. \end{aligned} \quad (\text{A.5.13})$$

Let us now lower-bound $f(p)$. In the same fashion as above, we have that

$$\begin{aligned} f(p) &= \int_{\mathcal{K} - \Sigma p} e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' \\ &\geq \int_{\overline{\mathbb{B}}(0, \tilde{R})} e^{-\frac{1}{2} x' \cdot \Sigma^{-1} x'} dx' \\ &\geq (2\pi)^{d/2} \det(\Sigma)^{1/2} \left(1 - 2^{d/2} e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} \right), \end{aligned} \quad (\text{A.5.14})$$

so that, if $\tilde{R} \geq \sqrt{\|\Sigma\|(2d+4)\log 2}$, it holds that

$$f(p) \geq \frac{1}{2} (2\pi)^{d/2} \det(\Sigma)^{1/2} \quad (\text{A.5.15})$$

The Hessian of $\log f/f(0)$ is then given by

$$\text{Hess } \log \frac{f(p)}{f(0)} = \frac{\text{Hess } f(p)}{f(p)} - \frac{\nabla f(p) \nabla f(p)^\top}{f(p)^2}. \quad (\text{A.5.16})$$

(A.5.11), (A.5.13), and (A.5.15) combined yield that

$$\begin{aligned} \left\| \text{Hess} \log \frac{f(p)}{f(0)} \right\| &\leq e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} (\text{tr} \Sigma + \|\Sigma\|) 2^{d/2+1} + \left(e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} \sqrt{\text{tr} \Sigma} 2^{d/2+1} \right)^2 \\ &\leq e^{-\frac{\tilde{R}^2}{4\|\Sigma\|}} (\text{tr} \Sigma + \|\Sigma\|) 2^{d+3} = \tilde{E}(\Sigma, \tilde{R}). \end{aligned} \quad (\text{A.5.17})$$

Taylor-Lagrange inequality now yields the full result since $\log f(0)/f(0) = 0$ and $\nabla \log f(0) = 0$. $\blacksquare$

Lemma A.5.2. Consider $X \sim \mathcal{N}(0, \sigma^2 I_d)$ a multivariate Gaussian distribution with $\sigma^2 > 0$. For $R > 0$, define the truncated Gaussian r.v. X_R by conditioning X to the ball $\overline{\mathbb{B}}(0, R)$. Define

$$\tilde{\mathcal{H}}(p) := \log \mathbb{E} \left[e^{\langle p, X_R \rangle} \right] \quad (\text{A.5.18a})$$

$$\tilde{\mathcal{L}}(p) := \tilde{\mathcal{H}}^*(p), \quad (\text{A.5.18b})$$

and

$$E(\sigma^2, R) := e^{-\frac{R^2}{16\sigma^2}} 2^{d+3}(d+1) \quad (\text{A.5.19})$$

and assume that $R > 0$ satisfies

$$R \geq 4\sigma \sqrt{(d+3) \log 2 + \log(d+1)}. \quad (\text{A.5.20})$$

Then, for any $p \in \mathbb{R}^d$ such that $\|p\| \leq \frac{R}{2\sigma^2}$, $v \in \mathbb{R}^d$ such that $\|v\| \leq \frac{R}{4}$, it holds that

$$\left(1 - E(\sigma^2, R)\right) \frac{\sigma^2 \|p\|^2}{2} \leq \tilde{\mathcal{H}}(p) \leq \left(1 + E(\sigma^2, R)\right) \frac{\sigma^2 \|p\|^2}{2} \quad (\text{A.5.21a})$$

$$\left(1 - 2E(\sigma^2, R)\right) \frac{\|v\|^2}{2\sigma^2} \leq \tilde{\mathcal{L}}(v) \leq \left(1 + 2E(\sigma^2, R)\right) \frac{\|v\|^2}{2\sigma^2}. \quad (\text{A.5.21b})$$

Proof. First, let us show that, for any $r > 0$, $\nabla \tilde{\mathcal{H}}(\overline{\mathbb{B}}(0, r))$ is an open ball centered at 0. Since the distribution of X_R is invariant by rotation, this set can be rewritten as

$$\nabla \tilde{\mathcal{H}}(\overline{\mathbb{B}}(0, r)) = \left\{ v \in \mathbb{R}^d : \|v\| \in \left\{ \|\nabla \tilde{\mathcal{H}}(p)\| : p \in \overline{\mathbb{B}}(0, r) \right\} \right\}. \quad (\text{A.5.22})$$

But, by connectedness of $\overline{\mathbb{B}}(0, r)$ and continuity of $\nabla \tilde{\mathcal{H}}$, $\left\{ \|\nabla \tilde{\mathcal{H}}(p)\| : p \in \overline{\mathbb{B}}(0, r) \right\}$ is an interval. Since $\nabla \tilde{\mathcal{H}}(0) = 0$, it is either $[0, \|\nabla \tilde{\mathcal{H}}(r)\|]$ or $[0, \|\nabla \tilde{\mathcal{H}}(r)\|]$. $\nabla \tilde{\mathcal{H}}(\overline{\mathbb{B}}(0, r))$ being compact and therefore closed, it must be the latter. Hence, we have shown that

$$\nabla \tilde{\mathcal{H}}(\overline{\mathbb{B}}(0, r)) = \overline{\mathbb{B}}\left(0, \sup_{\overline{\mathbb{B}}(0, r)} \|\nabla \tilde{\mathcal{H}}\|\right). \quad (\text{A.5.23})$$

We apply Lemma A.5.1 with $\tilde{R} \leftarrow \frac{R}{2}$. Note that our choice of R implies that $\tilde{R}$ satisfies the condition of Lemma A.5.1 and, moreover, that $E(\sigma^2, R) \leq \frac{1}{2}$. Lemma A.5.1 directly implies the bound on $\tilde{\mathcal{H}}$.

Consider $p \in \mathbb{R}^d$ such that $\|p\| \leq \frac{R}{2\sigma^2}$. Then, by Lemma A.5.1, we have that

$$\|\nabla \tilde{\mathcal{H}}(p)\| \geq \sigma^2 \|p\| - e^{-\frac{R^2}{16\sigma^2}} 2^{d+3} \left(\text{tr}(\sigma^2 I_d) + \|\sigma^2 I_d\| \right) \|p\|$$

$$\begin{aligned}
&= \sigma^2 \|p\| (1 - E(\sigma^2, R)) \\
&\geq \frac{\sigma^2 \|p\|}{2},
\end{aligned} \tag{A.5.24}$$

where we used that $E(\sigma^2, R) \leq \frac{1}{2}$. Therefore, we obtain that $\sup \left\{ \|\nabla \tilde{\mathcal{H}}(p)\| : p \in \overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right) \right\} \geq \frac{R}{4}$ so that $\nabla \tilde{\mathcal{H}}\left(\overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right)\right)$ contains $\overline{\mathbb{B}}\left(0, \frac{R}{4}\right)$.

Take $v \in \overline{\mathbb{B}}\left(0, \frac{R}{4}\right)$ which therefore belongs to $\nabla \tilde{\mathcal{H}}\left(\overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right)\right)$. Therefore, $\tilde{\mathcal{L}}(v)$ can be rewritten as

$$\tilde{\mathcal{L}}(v) = \sup_{p \in \overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right)} \langle v, p \rangle - \tilde{\mathcal{H}}(p). \tag{A.5.25}$$

Using [Lemma A.5.1](#) again, we obtain that

$$\sup_{p \in \overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right)} \langle v, p \rangle - \frac{\sigma^2}{2} (1 + E(\sigma^2, R)) \|v\|^2 \leq \tilde{\mathcal{L}}(v) \leq \sup_{p \in \overline{\mathbb{B}}\left(0, \frac{R}{2\sigma^2}\right)} \langle v, p \rangle - \frac{\sigma^2}{2} (1 - E(\sigma^2, R)) \|v\|^2, \tag{A.5.26}$$

so that, since $E(\sigma^2, R) \leq \frac{1}{2}$, we obtain that

$$\frac{\|v\|^2}{2\sigma^2(1 + E(\sigma^2, R))} \leq \tilde{\mathcal{L}}(v) \leq \frac{\|v\|^2}{2\sigma^2(1 - E(\sigma^2, R))}. \tag{A.5.27}$$

Since, for $x \in [0, 1/2]$, both $\frac{1}{1+x} \geq 1 - 2x$ and $\frac{1}{1-x} \leq 1 + 2x$ hold, we get

$$\frac{\|v\|^2}{2\sigma^2} (1 - 2E(\sigma^2, R)) \leq \tilde{\mathcal{L}}(v) \leq \frac{\|v\|^2}{2\sigma^2} (1 + 2E(\sigma^2, R)) \tag{A.5.28}$$

which concludes the proof. $\blacksquare$

We will require the following technical lemma.

Lemma A.5.3. Consider $v, w \in \mathbb{R}^d$ such that $0 < \|w\| \leq \frac{\mu R}{2}$ for some $R, \mu > 0$. Define,

$$f(u) = \sup_{p \in \mathbb{R}^d: \|p\| \leq R} \langle p, u \rangle - \frac{\mu}{2} \|p\|^2, \tag{A.5.29}$$

then, with $\lambda = \frac{\|v\|}{\|w\|}$,

$$\lambda f\left(\frac{v}{\lambda} + w\right) \leq f(v + w). \tag{A.5.30}$$

Proof. Define $p := \frac{1}{\mu} \left(\frac{v}{\lambda} + w\right)$ which has norm at most R since $\|w\| \leq \frac{R}{2\mu}$. Then, we have that

$$\begin{aligned}
\lambda f\left(\frac{v}{\lambda} + w\right) - f(v + w) &\leq \lambda \left(\left\langle p, \frac{v}{\lambda} + w \right\rangle - \frac{\mu}{2} \|p\|^2 \right) - \left(\langle p, v + w \rangle - \frac{\mu}{2} \|p\|^2 \right) \\
&= (\lambda - 1) \left(\langle p, w \rangle - \frac{\mu}{2} \|p\|^2 \right) \\
&= (\lambda - 1) \times \frac{1}{\mu} \left(\frac{\langle v, w \rangle}{\lambda} + \|w\|^2 - \frac{1}{2} \left(2\|w\|^2 + 2\frac{\langle v, w \rangle}{\lambda} \right) \right) \\
&= 0,
\end{aligned} \tag{A.5.31}$$

and our proof is complete. $\blacksquare$

B

THE GLOBAL CONVERGENCE TIME OF STOCHASTIC GRADIENT DESCENT IN NON-CONVEX LANDSCAPES: SHARP ESTIMATES VIA LARGE DEVIATIONS

This appendix contains the proofs for [Chapter 5](#). It is extracted from [Azizian et al. \(2025a\)](#).

B.1 LARGE DEVIATION ANALYSIS OF SGD

B.1.1 *Outline of proof*

Before reviewing notations and assumptions, we provide an overview of the proof of our main results [Theorem 5.7](#) and [Theorem 5.10](#).

Our framework is built on a large deviation principle for SGD ([Appendix B.1.4](#)): it yields precise estimations of the probability that SGD approximately follows an arbitrary continuous-time path. In particular, this LDP concerns the accelerated (subsamped) SGD process x_n^η , $n = 0, 1, \dots$, defined as

$$x_n^\eta = x_{n\lfloor 1/\eta \rfloor} \quad (\text{B.1.1})$$

We rely on the LDP from ([Azizian et al., 2024b](#)) that we restate in [Appendix B.1.4](#). Note that this LDP is obtained as a consequence of ([Freidlin and Wentzell, 1998b](#), Chap. 7) by comparing the process $(x_n^\eta)_{n \geq 0}$ to its continuous-time linear interpolation $(X(t))_{t \geq 0}$ defined by

$$X(t) = x_n + \frac{t - n\eta}{\eta}(x_{n+1} - x_n) \quad (\text{B.1.2})$$

for all $n = 0, 1, \dots$, and all $t \in [n\eta, (n+1)\eta]$. This is why the action function ([B.1.17](#)) is defined in terms of continuous time.

Our framework also builds on two other key results from ([Azizian et al., 2024b](#)). First, we need [Lemma B.1.4](#) restated in [Appendix B.1.5](#) that provides a control on the return time of SGD to a neighborhood of the set of critical points. In particular, it shows that the distribution of this return time is roughly sub-exponential. Second, we require the careful analysis done in ([Azizian et al., 2024b](#)) that exploits the structure of the set of critical points to obtain regularity properties on ([B.1.17](#)) locally, see [Appendix B.1.6](#).

In [Appendix B.2](#), we study the transitions of SGD, or more precisely the sequence of ([B.1.1](#)), between connected components of critical points. In short, we characterize the probability that the iterates of ([B.1.1](#)) transition from one component of critical points to another as well as the time it takes to do so. It is because of this new requirement compared to ([Azizian et al., 2024b](#)) that we have to refine their estimates. For this, we introduce an induced Markov

chain $(z_n)_{n \geq 0}$ (Definition B.12) that lives on the set of critical components and captures the essential global convergence properties.

With the transition probabilities and times in hand, the system can now be analyzed as a finite state space Markov chain whose states are the components of the set of critical points. Leveraging the tools of Freidlin and Wentzell (1998b), restated in Appendix B.3.1, we obtain a characterization of the global convergence time of (B.1.1) to the set of critical points in Appendix B.3.2.

The only thing left to do is to translate the results obtained for the subsampled process of (B.1.1) back to the original process of SGD (Appendix B.3.3). The upper-bound is immediate but the lower-bound needs some care. This is where the attracting strength defined in Appendix B.1.2 comes into play: if the attracting strength is large enough, we can show that when SGD reaches a neighborhood of a global minimum, it will remain there enough time for the next iterate of the subsampled process of (B.1.1) to still be in that neighborhood. This in turn allows us to show the global convergence times of SGD and the subsampled sequence roughly coincide, concluding the proof of Theorem 5.7.

Appendix B.3.4 is then dedicated to obtaining Theorem 5.10 while Appendix B.3.6 establishes the link between the presence of spurious local minima and a non-zero energy.

The results of Appendix B.4 are then obtained in Appendix B.4.

For convenience, the main dependencies between sections of the appendix are summarized in Figure 26.

B.1.2 Setup and assumptions

Before we begin our proof, we introduce here notations and we revisit and discuss our standing assumptions. In particular, to extend the range of our results, we provide in the rest of this appendix a weaker version of the blanket assumptions of Appendix B.1.2 which will be in force throughout the appendix.

We equip $\mathbb{R}^d$ with the canonical inner product $\langle \cdot, \cdot \rangle$ and the associated Euclidean norm $\|\cdot\|$. We denote by $\mathbb{B}(x, r)$ (resp. $\overline{\mathbb{B}}(x, r)$) the open (resp. closed) ball of radius r centered at x .

We also define, for any $A \subset \mathbb{R}^d$,

$$\mathcal{U}_\delta(A) := \{x \in \mathbb{R}^d : d(x, A) < \delta\}. \quad (\text{B.1.3})$$

For any $T > 0$, we denote by $\mathcal{C}([0, T]) = \mathcal{C}([0, T], \mathbb{R}^d)$ the set of continuous functions from $[0, T]$ to $\mathbb{R}^d$.

We begin with our assumptions for the objective function f , which are a weaker version of Assumption B.1.

Assumption B.1. The objective function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the following conditions:

- (a) *Coercivity:* $f(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$.
- (b) *Smoothness:* f is C^2 -differentiable and its gradient is β -Lipschitz continuous, namely

$$\|\nabla f(x') - \nabla f(x)\| \leq \beta \|x' - x\| \quad \text{for all } x, x' \in \mathbb{R}^d. \quad (\text{B.1.4})$$

Assumption B.2 (Critical set regularity). The critical set

$$\text{crit}(f) := \{x \in \mathbb{R}^d : \nabla f(x) = 0\} \quad (\text{B.1.5})$$

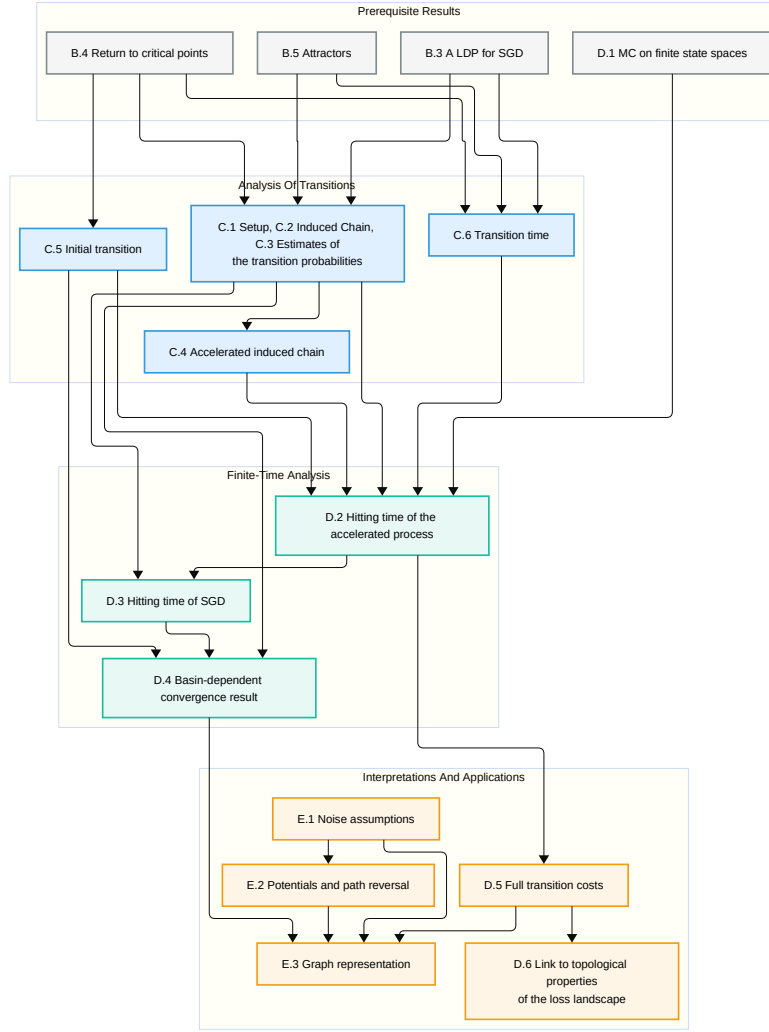


Figure 26: Structure of the proof: arrows represent the main logical implications between the main parts of the proof. The group “Prerequisite results” gathers the tools from the literature that we rely on, mostly either from [4, 20, 21]. Our key technical contributions that lead to Thm. 1-2 are “Analysis Of Transitions” and “Finite Time Analysis”. The results of §5 come from “Interpretation and Applications”.

of f consists of a finite number of (compact) connected components. Moreover, each of these components $\mathcal{K}$ is connected by piecewise absolutely continuous paths, i.e., for any $x, x' \in \mathcal{K}$, there exists $\gamma \in \mathcal{C}([0, 1], \mathcal{K})$ such that $\gamma_0 = x, \gamma_1 = x'$ and such that it is piecewise absolutely continuous, i.e., γ is differentiable almost everywhere and there exists $0 = t_0 < t_1 < \dots < t_N = 1$ such that $\dot{\gamma}$ is integrable on every closed interval of (t_n, t_{n+1}) for $n = 0, \dots, N - 1$.

Note that, since connected components of a closed set are closed, [Assumption B.1](#) automatically ensure that the connected components of $\text{crit}(f)$ are compact.

Unlike in [Assumption B.1](#), we do not require the connected components of $\text{crit } f$ to be smoothly connected but only piecewise absolutely continuous.

Remark 5. The path-connectedness requirement of [Assumption B.2](#) is satisfied whenever the connected components of $\text{crit}(f)$ are isolated critical points,

smooth manifolds, or finite unions of closed manifolds. More generally, [Assumption B.2](#) is satisfied whenever f is definable – in which case $\text{crit } f$ is also definable, so each component can be connected by piecewise smooth paths ([Coste, 1999](#), [Van Den Dries and Miller, 1996](#)). The relaxation provided by [Assumption B.2](#) represents the “minimal” set of hypotheses that are required for our analysis to go through. 

Moving forward, to align our notation with standard conventions in large deviations theory, it will be more convenient to work with $-Z(x; \omega)$ instead of $Z(x; \omega)$ in our proofs. To make this clear, we restate below [Assumption B.3](#) in terms of the noise process

$$U(x, \omega) = -Z(x; \omega). \quad (\text{B.1.6})$$

We also take the chance to relax the definition of the variance proxy of U , which requires the new assumption [Assumption B.4](#).

Assumption B.3. The error term $U: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}^d$ satisfies the following properties:

- (a) *Properness:* $\mathbb{E}[U(x, \omega)] = 0$ and $\text{cov}(U(x, \omega)) > 0$ for all $x \in \mathbb{R}^d$.
- (b) *Smooth growth:* $U(x, \omega)$ is C^2 -differentiable and satisfies the growth condition

$$\sup_{x, \omega} \frac{\|U(x, \omega)\|}{1 + \|x\|} < +\infty. \quad (\text{B.1.7})$$

- (c) *Sub-Gaussian tails:* There is $\sigma_\infty^2: \mathbb{R} \rightarrow (0, +\infty)$ continuous, with $\inf_{\mathbb{R}^d} \sigma_\infty^2 > 0$, such that $U(x, \omega)$ satisfies

$$\log \mathbb{E} \left[e^{\langle p, U(x, \omega) \rangle} \right] \leq \frac{\sigma_\infty^2(f(x))}{2} \|p\|^2 \quad \text{for all } p \in \mathbb{R}^d. \quad (\text{B.1.8})$$

Assumption B.4. The signal-to-noise ratio of G is bounded as

$$\liminf_{\|x\| \rightarrow \infty} \frac{\|\nabla f(x)\|^2}{\sigma_\infty^2 \circ f(x)} > 16 \log 6 \cdot d. \quad (\text{B.1.9})$$

Furthermore, $\sigma_\infty^2(t) = o(t^2)$ as $t \rightarrow +\infty$ and, $\frac{\sigma_\infty^2 \circ f(x)}{\|x\|^s}$ is bounded above and below at infinity for $s \in [0, 2]$, i.e.,

$$0 < \liminf_{\|x\| \rightarrow +\infty} \frac{\sigma_\infty^2 \circ f(x)}{\|x\|^s} \quad \text{and} \quad \limsup_{\|x\| \rightarrow +\infty} \frac{\sigma_\infty^2 \circ f(x)}{\|x\|^s} < +\infty. \quad (\text{B.1.10})$$

Remark 6. The key distinction between [Assumption B.3](#) and [Assumptions B.3–B.4](#) stems from their differing requirements for the variance proxy σ_∞^2 of the noise in [\(B.1.12\)](#). Since f is coercive, allowing σ_∞^2 to depend on $f(x)$ enables us to consider noise processes whose variance may grow unbounded as $\|x\| \rightarrow \infty$; we specifically choose to express this dependence through $f(x)$ rather than directly through x as it substantially simplifies both our proofs and calculations. A version of [Assumption B.23](#) adapted to this generalized setting will be stated as [Assumption B.23](#) in [Appendix B.4.1](#). 

We now introduce the notation for the target set $\mathcal{Q}$. In the main text, we choose $\mathcal{Q}$ to be the set of global minima of f but this needs not be the case in general. In the rest of the appendix, $\mathcal{Q}$ will be a union of connected components of $\text{crit}(f)$.

Definition B.5 (Choice of the components of interest). Denote by $\mathcal{Q}_1, \dots, \mathcal{Q}_{N_{\text{targ}}}$ the connected components of $\text{crit}(f)$ that form the *target* set.

Denote by $\mathcal{K}_1, \dots, \mathcal{K}_K$ the remaining connected components of $\text{crit}(f)$ and denote by

$$(\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}) := (\mathcal{K}_1, \dots, \mathcal{K}_K, \mathcal{Q}_1, \dots, \mathcal{Q}_{N_{\text{targ}}}) \quad (\text{B.1.11})$$

their union with $N_{\text{crit}} = K + N_{\text{targ}}$.

Since the connected components of $\text{crit}(f)$ are pairwise disjoint by definition, their compactness implies that there exists $\delta > 0$ such that $\mathcal{U}_\delta(\mathcal{C}_i)$, for $i = 1, \dots, N_{\text{crit}}$, are pairwise disjoint.

In this framework, the iterates of (B.1.12), started at $x \in \mathbb{R}^d$, are defined by the following recursion:

$$\begin{cases} x_0 \in \mathbb{R}^d \\ x_{n+1} = x_n - \eta \nabla f(x_n) + \eta U_n, \quad \text{where } U_n = U(x_n, \omega_n) \end{cases} \quad (\text{B.1.12})$$

where $(\omega_n)_{n \geq 0}$ is a sequence of random variables in $\mathbb{R}^m$. We will denote by $\mathbb{P}_x$ the law of the sequence $(\omega_n)_{n \geq 0}$ when the initial point is x and by $\mathbb{E}_x$ the expectation with respect to $\mathbb{P}_x$.

[Assumption B.1](#) and [Assumption B.3](#) imply the following growth condition, that we assume holds with the same constant for the sake of simplicity. There is $M > 0$ such that, for all $x \in \mathbb{R}^d$, $\omega \in \Omega$,

$$\|\nabla f(x)\| \leq M(1 + \|x\|) \quad \text{and} \quad \|U(x, \omega)\| \leq M(1 + \|x\|). \quad (\text{B.1.13})$$

B.1.3 Hamiltonian and Lagrangian

Following the notation of ([Azizian et al., 2024b](#)), we introduce the cumulant generating functions of the noise $U(x, \omega)$ and of the drift $-\nabla f(x) + U(x, \omega)$, that we denote by $\tilde{\mathcal{H}}, \mathcal{H}$ to avoid confusion. We also define their convex conjugates, $\tilde{\mathcal{L}}, \mathcal{L}$.

Definition B.6 (Hamiltonian and Lagrangian). Define, for $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$,

$$\tilde{\mathcal{H}}(x, v) = \log \mathbb{E}[\exp(\langle v, U(x, \omega) \rangle)] \quad (\text{B.1.14a})$$

$$\mathcal{H}(x, v) = -\langle \nabla f(x), v \rangle + \tilde{\mathcal{H}}(x, v) \quad (\text{B.1.14b})$$

$$\tilde{\mathcal{L}}(x, v) = \tilde{\mathcal{H}}(x, \cdot)^*(v) \quad (\text{B.1.14c})$$

$$\mathcal{L}(x, v) = \mathcal{H}(x, \cdot)^*(v) = \tilde{\mathcal{L}}(x, v + \nabla f(x)). \quad (\text{B.1.14d})$$

$\mathcal{L}$ is thus equal to the Lagrangian $\mathcal{L}_G(\cdot, \cdot)$.

We restate ([Azizian et al., 2024b](#), Lem. B.1) that provides basic properties of the Hamiltonian and Lagrangian functions.

Lemma B.1.1 (Properties of $\mathcal{H}$ and $\mathcal{L}$, ([Azizian et al., 2024b](#), Lem. B.1)).

1. $\mathcal{H}$ is $\mathcal{C}^2$ and $\mathcal{H}(x, \cdot)$ is convex for any $x \in \mathbb{R}^d$.
2. $\mathcal{L}(x, \cdot)$ is convex for any $x \in \mathbb{R}^d$, $\mathcal{L}$ is l.s.c. on $\mathbb{R}^d \times \mathbb{R}^d$.
3. For any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$, $\mathcal{H}(x, v) \leq 2M(1 + \|x\|)\|v\|^2$ and $\text{dom } \mathcal{L}(x, \cdot) \subset \overline{\mathbb{B}}(0, 2M(1 + \|x\|))$.
4. For any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$, $\mathcal{L}(x, v) \geq 0$ and $\mathcal{L}(x, v) = 0 \iff v = \nabla f(x)$.

The following lemma provides a lower bound on the Lagrangian and is an immediate consequence of the sub-Gaussian tails assumption ((c)).

Lemma B.1.2 ((Azizian et al., 2024b, Lem. D.5)). For any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$,

$$\mathcal{L}(x, v) \geq \frac{\|v + \nabla f(x)\|^2}{2\sigma_\infty^2 \circ f(x)}. \quad (\text{B.1.15})$$

B.1.4 A Large Deviations Principle for SGD

In this section, we present and restate the large deviation principles established in (Azizian et al., 2024b) for SGD. Note that their proof is itself an application of the general theory of Freidlin and Wentzell (1998b). From the sequences $(x_n)_{n \geq 0}$ and $(\omega_n)_{n \geq 0}$, we define another discrete sequence: a subsampled or, accelerated, sequence

$$x_n^\eta := x_{n \lfloor 1/\eta \rfloor}. \quad (\text{B.1.16})$$

From the Lagrangian defined in (B.1.14d), we define, on $\mathcal{C}([0, T], \mathbb{R}^d)$, the normalized action functional $\mathcal{S}_{0,T}$ by

$$\mathcal{S}_{0,T}(\gamma) = \begin{cases} \int_0^T \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt & \text{if } \gamma \text{ absolutely continuous} \\ +\infty & \text{otherwise} \end{cases} \quad (\text{B.1.17})$$

following Freidlin and Wentzell (1998b, Chap. 3.2), as a manner to quantify how “probable” a trajectory is.

For some $N > 0$, we will first equip $(\mathbb{R}^d)^N$ with the distance

$$\text{dist}_N(\xi, \zeta) = \max_{0 \leq n \leq N-1} \|\xi_n - \zeta_n\|. \quad (\text{B.1.18})$$

Now, for $N \geq 0$, $\xi = (\xi_0, \dots, \xi_{N-1}) \in (\mathbb{R}^d)^N$, let us define the normalized discrete action functional

$$\mathcal{S}_N(\xi) := \sum_{n=0}^{N-2} \rho(\xi_n, \xi_{n+1}) \quad (\text{B.1.19})$$

where the cost of moving from one iteration to the next is defined for any $x, x' \in \mathbb{R}^d$ from the previous continuous normalized action functional (cf. (B.1.17)) with horizon 1 as

$$\rho(x, x') := \inf\{\mathcal{S}_{0,1}(\gamma) : \gamma \in \mathcal{C}([0, 1], \mathbb{R}^d), \gamma_0 = x, \gamma_1 = x'\}. \quad (\text{B.1.20})$$

We now present the large deviation principle on the discrete accelerated sequence $(x_n^\eta)_{0 \leq n \leq N-1} = (x_{n \lfloor 1/\eta \rfloor})_{0 \leq n \leq N-1}$. In the following result, the functional $\eta^{-1} \mathcal{S}_N$ is thus the action functional in $(\mathbb{R}^d)^N$ of the process $(x_n^\eta)_{0 \leq n \leq N-1}$ uniformly with respect to the starting point x_0 in any compact set $\mathcal{K} \subset \mathbb{R}^d$, as $\eta \rightarrow 0$.

Proposition B.1.3. Fix $N \geq 0$.

- For any $s > 0$, the set

$$\Gamma_N^{\mathcal{K}}(s) := \{\xi \in (\mathbb{R}^d)^N : \xi_0 \in \mathcal{K}, \mathcal{S}_N(\xi) \leq s\} \quad (\text{B.1.21})$$

is compact and $\mathcal{S}_N$ is l.s.c. on $(\mathbb{R}^d)^N$.

- For any $s, \delta, \varepsilon > 0$, $\mathcal{K} \subset \mathbb{R}^d$ compact, there exists $\eta_0 > 0$ such that, for any $\eta \in (0, \eta_0]$, for any $x_0 \in \mathcal{K}$, $n \leq N$, $\xi \in \Gamma_n^{\{x_0\}}(s)$, we have that

$$\mathbb{P}_{x_0}(\text{dist}_n(x^\eta, \xi) < \delta) \geq \exp\left(-\frac{\mathcal{S}_n(\xi) + \varepsilon}{\eta}\right) \quad (\text{B.1.22a})$$

$$\mathbb{P}_{x_0}(\text{dist}_n(x^\eta, \Gamma_n^{\{x_0\}}(s)) > \delta) \leq \exp\left(-\frac{s - \varepsilon}{\eta}\right). \quad (\text{B.1.22b})$$

B.1.5 Return to critical points

In this section, we restate a key result from (Azizian et al., 2024b): the lemma below provides a control on the return time to a neighborhood of the set of critical points. In particular, it shows that the distribution of this return time is roughly sub-exponential.

Definition B.7 (Stopping times for the accelerated process). For any set $A \subset \mathbb{R}^d$, we define the hitting and exit times of A :

$$\tau_A^\eta := \inf\{n \geq 0 : x_n^\eta \in A\}, \quad (\text{B.1.23a})$$

$$\sigma_A^\eta := \inf\{n \geq 0 : x_n^\eta \notin A\}. \quad (\text{B.1.23b})$$

Lemma B.1.4 ((Azizian et al., 2024b, Lem. D.21)). Consider $\text{crit}(f) \subset \mathcal{U} \subset \mathcal{X} \subset \mathbb{R}^d$ with $\mathcal{U}$ an open set and $\mathcal{X}$ a compact set. Then, there is some $\eta_0, \alpha_0, a, b > 0$ such that,

$$\forall \eta \leq \eta_0, \alpha \leq \alpha_0, x \in \mathcal{X}, \quad \mathbb{E}_x \left[e^{\frac{\alpha \tau_{\mathcal{U}}^\eta}{\eta}} \right] \leq e^{\frac{a\alpha}{\eta} + b}. \quad (\text{B.1.24})$$

B.1.6 Attractors

We again build on the work of (Azizian et al., 2024b) which took inspiration from the framework of Kifer (1988b). We first need to define the gradient flow of f .

Definition B.8. Define, for $x \in \mathbb{R}^d$, the flow Θ of $-\nabla f$ started at x , i.e.,

$$\Theta_0(x) = x \quad (\text{B.1.25})$$

$$\dot{\Theta}_t(x) = -\nabla f(\Theta_t(x)). \quad (\text{B.1.26})$$

and let $F(x)$ be the value of this flow at time 1, i.e.,

$$F(x) = \Theta_1(x). \quad (\text{B.1.27})$$

We first list some basic properties.

Lemma B.1.5 (Properties of the flow (Azizian et al., 2024b, Lem. D.1)). Θ is well-defined and continuous in both time and space, and, for any $T \geq 0$, $\gamma \in \mathcal{C}([0, T], \mathbb{R}^d)$ such that $\gamma_0 = x$,

$$\mathcal{S}_{0,T}(\gamma) = 0 \iff \gamma_t = \Theta_t(x) \text{ for all } t \in [0, T]. \quad (\text{B.1.28})$$

The following lemma translates this for F .

Lemma B.1.6 (Properties of F , (Azizian et al., 2024b, Lem. D.2)). F is well-defined and continuous and, for any $x, x' \in \mathbb{R}^d$,

$$\rho(x, x') = 0 \iff x' = F(x). \quad (\text{B.1.29})$$

The next two lemmas are key regularity results on the connected components of the critical set.

Lemma B.1.7 ((Azizian et al., 2024b, Lem. D.8)). For any $\mathcal{C} \subset \text{crit}(f)$ connected component of the critical set, there is $r_0 > 0$ such that, for any $0 < r \leq r_0$,

$$\mathcal{W}_r(\mathcal{C}) := \{x \in \mathbb{R}^d : \rho(x, \mathcal{C}) < r, \rho(\mathcal{C}, x) < r\} \quad (\text{B.1.30})$$

is open and contains $\mathcal{C}$.

Lemma B.1.8 ((Azizian et al., 2024b, Lem. D.9)). Let $\mathcal{C} \subset \text{crit}(f)$ be a connected component of the critical set. Then, for any $\varepsilon > 0$, there is some $N \geq 1$ such that, for any $x, z \in \mathcal{C}$, there is $\xi \in \mathcal{D}_r(N)$ such that $\xi_0 = x$, $\xi_{N-1} = z$, $\mathcal{S}_N(\xi) < \varepsilon$ and $\max_{0 \leq n < N} d(\xi_n, \mathcal{C}) < \varepsilon$.

B.1.7 Convergence and stability

Lemma B.1.9 ((Azizian et al., 2024b, Lem. D.28)). For any $x \in \mathcal{X}_r$, there exists $i \in \{1, \dots, N_{\text{crit}}\}$ such that

$$\lim_{t \rightarrow +\infty} d(\Theta_t(x), \mathcal{C}_i) = 0. \quad (\text{B.1.31})$$

Definition B.9. A connected component of the critical points $\mathcal{C} \subset \text{crit}(f)$ is said to be *asymptotically stable* if there exists $\mathcal{U}$ a neighborhood of $\mathcal{C}$ such that, for any $x \in \mathcal{U}$, $\Theta_t(x)$ converges to $\mathcal{C}$, i.e.,

$$\lim_{t \rightarrow +\infty} d(\Theta_t(x), \mathcal{C}) = 0. \quad (\text{B.1.32})$$

The notions of minimizing component and asymptotic stability are equivalent in our context.

Definition B.10. $\mathcal{C}$ connected component of $\text{crit}(f)$ is *minimizing* if there exists $\mathcal{U}$ a neighborhood of $\mathcal{C}$ such that

$$\arg \min_{x \in \mathcal{U}} f(x) = \mathcal{C} \quad (\text{B.1.33})$$

Note that since $\text{crit}(f)$ is closed, $\mathcal{C}$ is closed as well as a connected component of a closed set.

Lemma B.1.10 ((Azizian et al., 2024b, Lem. D.29)). For any $\mathcal{C} \subset \text{crit}(f)$ connected component of the set of critical points, $\mathcal{C}$ is minimizing if and only if it is asymptotically stable.

B.2 TRANSITIONS

In this section, we study the transitions of the accelerated sequence of SGD iterates between the different sets of critical points. As in (Azizian et al., 2024b), we build on the work of Kifer (1988b) and Freidlin and Wentzell (1998b). More precisely, Appendix B.2.1, Appendix B.2.2, Appendix B.2.3, Appendix B.2.4, and Appendix B.2.5 consists in refining the framework of (Azizian et al., 2024b) to

be able to obtain precise time estimates on the transitions between the different sets of critical points. Such results are provided in [Appendix B.2.6](#).

B.2.1 Setup

We adapt to our context [Kifer \(1988b, Lem. 5.4\)](#) and simplify it using ideas from [Freidlin and Wentzell \(1998b, Chap. 6\)](#).

Definition B.11 ([Freidlin and Wentzell \(1998b, Chap. 6, §2\)](#)). For $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, $\delta > 0$,

$$\begin{aligned} \mathcal{B}_{i,j} &:= \inf \left\{ \mathcal{S}_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 \in \mathcal{C}_i, \xi_{N-1} \in \mathcal{C}_j, \xi_n \notin \bigcup_{l \neq i,j} \mathcal{C}_l \text{ for all } n = 1, \dots, N-2 \right\} \\ \mathcal{B}_{i,j}^\delta &:= \inf \left\{ \mathcal{S}_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 \in \mathcal{U}_\delta(\mathcal{C}_i), \xi_{N-1} \in \mathcal{U}_\delta(\mathcal{C}_j), \xi_n \notin \bigcup_{l \neq i,j} \mathcal{C}_l \text{ for all } n = 1, \dots, N-2 \right\} \end{aligned} \quad (\text{B.2.1})$$

While $\mathcal{B}_{i,j}$ is the usual definition from [Freidlin and Wentzell \(1998b, Chap. 6, §2\)](#), $\mathcal{B}_{i,j}^\delta$ is a variant that will prove helpful.

Let us now first list a few immediate properties.

Lemma B.2.1. For $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, $\mathcal{B}_{i,j}^\delta$ is non-decreasing in δ and, for any $\delta > 0$, $\mathcal{B}_{i,j}^\delta \leq \mathcal{B}_{i,j}$.

Note that the fact that $\mathcal{B}_{i,j}^\delta$ is non-decreasing in δ implies that the limit $\lim_{\delta \rightarrow 0} \mathcal{B}_{i,j}^\delta$ exists.

We now exploit the regularity around the $\mathcal{K}_i$, to obtain an alternate expression for $\mathcal{B}_{i,j}$ as the limit of $\mathcal{B}_{i,j}^\delta$ as $\delta \rightarrow 0$.

Lemma B.2.2. For $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, the following equality holds:

$$\mathcal{B}_{i,j} = \lim_{\delta \rightarrow 0} \mathcal{B}_{i,j}^\delta. \quad (\text{B.2.2})$$

Proof. By [Lemma B.2.1](#), we have

$$\mathcal{B}_{i,j}^\delta \leq \mathcal{B}_{i,j}, \quad (\text{B.2.3})$$

and therefore

$$\lim_{\delta \rightarrow 0} \mathcal{B}_{i,j}^\delta \leq \mathcal{B}_{i,j}. \quad (\text{B.2.4})$$

It remains to show the reverse inequality.

Take $\varepsilon > 0$. Apply [Lemma B.1.7](#) to $\mathcal{C}_i$: at the potential cost of reducing ε , we get that $\mathcal{W}_\varepsilon(\mathcal{C}_i)$ is an open neighborhood of $\mathcal{C}_i$.

Take $\delta > 0$ small enough so that $\mathcal{U}_\delta(\mathcal{C}_i) \subset \mathcal{W}_\varepsilon(\mathcal{C}_i)$, $\mathcal{U}_\delta(\mathcal{C}_i)$ does not intersect with any other $\mathcal{C}_l$, $l \neq j$.

Now, take any $N \geq 1$ and $\xi \in \mathcal{D}_r(N)$ such that $\xi_0 \in \mathcal{U}_\delta(\mathcal{C}_i)$, $\xi_{N-1} \in \mathcal{U}_\delta(\mathcal{C}_j)$, $\xi_n \notin \bigcup_{l \neq i,j} \mathcal{C}_l$ for all $n = 1, \dots, N-2$.

By the choice of δ , ξ_0 cannot be in $\bigcup_{l \neq i,j} \mathcal{C}_l$ and ξ_0 is inside $\mathcal{W}_\varepsilon(\mathcal{C}_i)$ so that there exists $x \in \mathcal{C}_i$ such that $\rho(x, \xi_0) < \varepsilon$. Similarly, there exists $x' \in \mathcal{C}_j$ such that $\rho(\xi_{N-1}, x') < \varepsilon$. Now consider the path $\zeta \in \mathcal{D}_r(N+2)$ defined as $\zeta = (x, \xi_0, \xi_1, \dots, \xi_{N-1}, x')$ that satisfies $\zeta_0 \in \mathcal{C}_i$, $\zeta_N \in \mathcal{C}_j$ and $\zeta_n \notin \bigcup_{l \neq i,j} \mathcal{C}_l$ for all

$n = 1, \dots, N-2$. Furthermore, $\mathcal{S}_{N+2}(\zeta) \leq \mathcal{S}_N(\xi) + 2\varepsilon$ and thus $\mathcal{B}_{i,j} \leq \mathcal{S}_N(\xi) + 2\varepsilon$. Passing to the infimum over such paths ξ yields

$$\mathcal{B}_{i,j}^\delta \leq \mathcal{B}_{i,j}^\delta + 2\varepsilon, \quad (\text{B.2.5})$$

which concludes the proof. $\blacksquare$

B.2.2 Induced chain

We now define an important object: the law of the (accelerated) iterated at the first time they reach some set $\mathcal{C}^\delta$ (typically a neighborhood of the critical set). Due to our interest in the finite-time dynamics, we slightly deviate from the classical definitions of Douc et al. (2018, Chap. 3.4) and Kifer (1988b, Prop. 5.3) to follow more closely the one of Freidlin and Wentzell (1998b, Chap. 6,§2).

Definition B.12. Consider $\mathcal{C}_i^\delta, i = 1, \dots, N_{\text{crit}}$ disjoint neighborhoods of $\mathcal{C}_i$ and denote by $\mathcal{C}^\delta := \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$ their union. We define recursively the sequences of stopping times $(\sigma_n^\eta)_{n \geq 0}$ and $(\tau_n^\eta)_{n \geq 0}$ by

- For $n = 0$,

$$\sigma_0^\eta := 0 \quad (\text{B.2.6})$$

$$\tau_0^\eta := 0. \quad (\text{B.2.7})$$

- For $n \geq 0$, if $x_{\tau_n^\eta}^\eta$ is in $\mathcal{C}_{i_n}^\delta$, then

$$\sigma_{n+1}^\eta := \begin{cases} \inf\{k \geq \tau_n^\eta : x_k^\eta \notin \mathcal{C}_{i_n}^\delta\} & \text{if } x_{\tau_n^\eta}^\eta \in \mathcal{C}_{i_n}^\delta \text{ for some } i_n \text{ s.t. } \mathcal{C}_{i_n} \in \{\mathcal{K}_1, \dots, \mathcal{K}_K\}, \\ \tau_n^\eta & \text{otherwise.} \end{cases} \quad (\text{B.2.8})$$

$$\tau_{n+1}^\eta := \inf\{k \geq \sigma_{n+1}^\eta : x_k^\eta \in \mathcal{C}^\delta\}. \quad (\text{B.2.9})$$

We denote by $z_n = x_{\tau_n^\eta}^\eta, n = 0, 1, \dots$, the induced Markov chain and by $\mathbb{Q}_\mathcal{C}^\delta(x, \cdot)$ the corresponding Markov transition probability i.e., the law of z_1 started at $z_0 = x$.

A few remarks are in order on $(z_n)_{n \geq 0}$:

- $z_n \in \mathcal{C}^\delta$ for all $n \geq 1$.
- If the chain reaches a neighborhood of $\mathcal{Q}_1, \dots, \mathcal{Q}_{N_{\text{targ}}}$ it stays there forever.

B.2.3 Estimates of the transition probabilities

Lemma B.2.3. For any $\varepsilon > 0, A > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i, i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such that for all $i \in \{1, \dots, K\}, j \in \{1, \dots, N_{\text{crit}}\}, x \in \mathcal{C}_i^\delta, 0 < \eta < \eta_0$,

$$\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) \leq \begin{cases} \exp\left(-\frac{\mathcal{B}_{i,j}}{\eta} + \frac{\varepsilon}{\eta}\right) & \text{if } \mathcal{B}_{i,j} < +\infty, \\ \exp\left(-\frac{A}{\eta}\right) & \text{otherwise.} \end{cases} \quad (\text{B.2.10})$$

Proof. Following the alternative definition of the $\mathcal{B}_{i,j}$'s [Lemma B.2.2](#), choose $\delta > 0$ small enough so that both $\mathcal{U}_{\frac{\delta}{2}}(\mathcal{C}_i)$, $i = 1, \dots, N_{\text{crit}}$ are pairwise disjoint and the following holds for all $i, j \in \{1, \dots, N_{\text{crit}}\}$:

$$\mathcal{B}_{i,j}^\delta \geq \begin{cases} \mathcal{B}_{i,j} - \varepsilon & \text{if } \mathcal{B}_{i,j} < +\infty, \\ A & \text{otherwise.} \end{cases} \quad (\text{B.2.11})$$

Require then that $\mathcal{C}_i^\delta$ be contained in $\mathcal{U}_{\frac{\delta}{2}}(\mathcal{C}_i)$ for all $i = 1, \dots, N_{\text{crit}}$. Note that the $\mathcal{C}_i^\delta$'s are pairwise disjoint by construction.

Given neighborhoods $\mathcal{C}_i^\delta$, $i = 1, \dots, N_{\text{crit}}$, by compactness of $\text{crit}(f)$, there exists $\delta' > 0$ such that $\delta' \leq \frac{\delta}{2}$, $\mathcal{U}_{2\delta'}(\mathcal{C}_i)$ is contained in $\mathcal{C}_i^\delta$ for all $i = 1, \dots, N_{\text{crit}}$.

Fix $i, j \in \{1, \dots, N_{\text{crit}}\}$ and consider $\xi \in \mathcal{D}_r(N)$ such that $\xi_0 \in \mathcal{U}_\delta(\mathcal{C}_i^\delta) \subset \mathcal{U}_\delta(\mathcal{C}_i)$, $\xi_{N-1} \in \mathcal{U}'_\delta(\mathcal{C}_j^\delta) \subset \mathcal{U}_\delta(\mathcal{C}_j)$ and $\xi_n \in \mathcal{U}_{\delta'}(\mathbb{R}^d \setminus \bigcup_{l \neq i,j} \mathcal{C}_l^\delta)$ for all $n = 1, \dots, N-2$. By the choice of δ' , ξ_n cannot be in $\bigcup_{l \neq i,j} \mathcal{U}'_\delta(\mathcal{C}_l)$ for any $n = 1, \dots, N-2$.

By construction of ξ we thus have that

$$\begin{aligned} \mathcal{S}_N(\xi) &\geq \mathcal{B}_{i,j}^\delta \\ &\geq \begin{cases} \mathcal{B}_{i,j} - \varepsilon & \text{if } \mathcal{B}_{i,j} < +\infty, \\ A & \text{otherwise.} \end{cases} \end{aligned} \quad (\text{B.2.12})$$

where the last inequality follows from [\(B.2.11\)](#).

Fix $x \in \mathcal{C}_i^\delta$. Let us now bound the probability

$$\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) = \mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}}^\eta \in \mathcal{C}_j^\delta \right), \quad (\text{B.2.13})$$

and start with the case where $\mathcal{B}_{i,j} < +\infty$.

We have, for any $N \geq 0$,

$$\mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}}^\eta \in \mathcal{C}_i^\delta \right) \leq \mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}}^\eta \in \mathcal{C}_i^\delta, \tau_{\mathcal{C}^\delta}^\eta < N \right) + \mathbb{P}_x \left(\tau_{\mathcal{C}^\delta}^\eta \geq N \right). \quad (\text{B.2.14})$$

We first bound the second probability using [Lemma B.1.4](#) applied to $\mathcal{U} \leftarrow \mathcal{C}_i^\delta$. Take N such that $\alpha_0(a - N) + \eta_0 b \leq -\mathcal{B}_{i,j}$. Then, by Markov's inequality and [Lemma B.1.4](#), it holds that for all $\eta \leq \eta_0$

$$\begin{aligned} \mathbb{P}_x \left(\tau_{\mathcal{C}^\delta}^\eta \geq N \right) &\leq \mathbb{P}_x \left(\exp \left(\frac{\alpha_0 \tau_{\mathcal{C}^\delta}^\eta}{\eta} \right) \geq \exp \left(\frac{\alpha_0 N}{\eta} \right) \right) \\ &\leq \exp \left(\frac{\alpha_0(a - N)}{\eta} + b \right) \\ &\leq \exp \left(\frac{-\mathcal{B}_{i,j}}{\eta} \right). \end{aligned} \quad (\text{B.2.15})$$

We now bound the term $\mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}}^\eta \in \mathcal{C}_j^\delta, \tau_{\mathcal{C}^\delta}^\eta < N \right)$ for this choice of N .

For this, we show that $x_{\tau_{\mathcal{C}^\delta}}^\eta \in \mathcal{C}_j^\delta$ with $\tau_{\mathcal{C}^\delta}^\eta < N$ implies that

$$\text{dist}_N \left(x^\eta, \Gamma_N^{\{x\}}(\mathcal{B}_{i,j} - 2\varepsilon) \right) > \frac{\delta'}{2}. \quad (\text{B.2.16})$$

Indeed, on the event $x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta$ with $\tau_{\mathcal{C}^\delta}^\eta < N$, there is some $N' \leq N$ such that $\tau_{\mathcal{C}^\delta}^\eta = N' - 1$. If $\text{dist}_N(x^\eta, \Gamma_N^{\{x\}}(\mathcal{B}_{i,j} - 2\varepsilon)) > \frac{\delta'}{2}$ did not hold, this would mean that there exists $\xi \in (\mathbb{R}^d)^{N'}$ such that $\text{dist}_{N'}(x^\eta, \xi) < \delta'$, $\xi_0 = x$ and $\mathcal{S}_{N'}(\xi) \leq \mathcal{B}_{i,j} - 2\varepsilon$. In particular, ξ would also satisfy $\xi_{N'-1} \in \mathcal{U}_{\delta'}(\mathcal{C}_j^\delta)$, $\xi_n \in \mathcal{U}_{\delta'}(\mathbb{R}^d \setminus \mathcal{C}^\delta)$ for all $n = 1, \dots, N' - 2$ and, as a consequence, $\xi \in \mathcal{D}_r(N')$. This would be thus in direct contradiction of (B.2.12).

Therefore, we have that

$$\begin{aligned} \mathbb{P}_x\left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta, \tau_{\mathcal{C}^\delta}^\eta < N\right) &\leq \mathbb{P}_x\left(\text{dist}_N\left(x^\eta, \Gamma_N^{\{x\}}(\mathcal{B}_{i,j} - 2\varepsilon)\right) > \frac{\delta'}{2}\right) \\ &\leq \exp\left(-\frac{\mathcal{B}_{i,j} - 3\varepsilon}{\eta}\right), \end{aligned} \quad (\text{B.2.17})$$

by Proposition B.1.3.

Combining this bound with (B.2.15) yields

$$\mathbb{P}_x\left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta\right) \leq \exp\left(-\frac{\mathcal{B}_{i,j} - 3\varepsilon}{\eta}\right) + \exp\left(\frac{-\mathcal{B}_{i,j}}{\eta}\right), \quad (\text{B.2.18})$$

which concludes the proof when $\mathcal{B}_{i,j} < +\infty$.

Let us now examine the case where $\mathcal{B}_{i,j} = +\infty$. Again, we have, for any $N \geq 0$,

$$\mathbb{P}_x\left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_i^\delta\right) \leq \mathbb{P}_x\left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_i^\delta, \tau_{\mathcal{C}^\delta}^\eta < N\right) + \mathbb{P}_x\left(\tau_{\mathcal{C}^\delta}^\eta \geq N\right). \quad (\text{B.2.19})$$

We first bound the second probability using Lemma B.1.4 applied to $\mathcal{U} \leftarrow \mathcal{C}_i^\delta$. Take N such that $\alpha_0(a - N) + \eta_0 b \leq -A$. Then, by Markov's inequality and Lemma B.1.4, it holds that for all $\eta \leq \eta_0$

$$\begin{aligned} \mathbb{P}_x\left(\tau_{\mathcal{C}^\delta}^\eta \geq N\right) &\leq \mathbb{P}_x\left(\exp\left(\frac{\alpha_0 \tau_{\mathcal{C}^\delta}^\eta}{\eta}\right) \geq \exp\left(\frac{\alpha_0 N}{\eta}\right)\right) \\ &\leq \exp\left(\frac{\alpha_0(a - N)}{\eta} + b\right) \\ &\leq \exp\left(\frac{-A}{\eta}\right). \end{aligned} \quad (\text{B.2.20})$$

We now bound the term $\mathbb{P}_x\left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta, \tau_{\mathcal{C}^\delta}^\eta < N\right)$ for this choice of N .

For this, we show that $x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta$ with $\tau_{\mathcal{C}^\delta}^\eta < N$ implies that

$$\text{dist}_N\left(x^\eta, \Gamma_N^{\{x\}}(A - \varepsilon)\right) > \frac{\delta'}{2}. \quad (\text{B.2.21})$$

Indeed, on the event $x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta$ with $\tau_{\mathcal{C}^\delta}^\eta < N$, there is some $N' \leq N$ such that $\tau_{\mathcal{C}^\delta}^\eta = N' - 1$. If $\text{dist}_N\left(x^\eta, \Gamma_N^{\{x\}}(A - \varepsilon)\right) > \frac{\delta'}{2}$ did not hold, this would mean that there exists $\xi \in (\mathbb{R}^d)^{N'}$ such that $\text{dist}_{N'}(x^\eta, \xi) < \delta'$, $\xi_0 = x$ and $\mathcal{S}_{N'}(\xi) \leq A - \varepsilon$. In particular, ξ would also satisfy $\xi_{N'-1} \in \mathcal{U}_{\delta'}(\mathcal{C}_j^\delta)$, $\xi_n \in \mathcal{U}_{\delta'}(\mathbb{R}^d \setminus \mathcal{C}^\delta)$ for all $n = 1, \dots, N' - 2$ and, as a consequence, $\xi \in \mathcal{D}_r(N')$. This would be thus in direct contradiction of (B.2.12).

Therefore, we have that

$$\begin{aligned} \mathbb{P}_x \left(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta, \tau_{\mathcal{C}^\delta}^\eta < N \right) &\leq \mathbb{P}_x \left(\text{dist}_N \left(x^\eta, \Gamma_N^{\{x\}}(A - \varepsilon) \right) > \frac{\delta'}{2} \right) \\ &\leq \exp \left(-\frac{A - 2\varepsilon}{\eta} \right), \end{aligned} \quad (\text{B.2.22})$$

by [Proposition B.1.3](#).

Combining this bound with [\(B.2.20\)](#) yields

$$\mathbb{P}_x(x_{\tau_{\mathcal{C}^\delta}^\eta}^\eta \in \mathcal{C}_j^\delta) \leq \exp \left(-\frac{A - 2\varepsilon}{\eta} \right) + \exp \left(\frac{-A}{\eta} \right), \quad (\text{B.2.23})$$

which concludes the proof. $\blacksquare$

Lemma B.2.4. *For any $\varepsilon > 0$, for any neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$ small enough, there exists $\eta_0 > 0$ such that for all $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, $x \in \mathcal{C}_i^\delta$, $0 < \eta < \eta_0$,*

$$\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) \geq \exp \left(-\frac{\mathcal{B}_{i,j}}{\eta} - \frac{\varepsilon}{\eta} \right). \quad (\text{B.2.24})$$

Note that this result is trivially valid if $\mathcal{B}_{i,j} = +\infty$.

Proof. For any $(i, j) \in \{1, \dots, K\} \times \{1, \dots, N_{\text{crit}}\}$, there exists $N_{i,j} \geq 1$, $\xi^{i,j} \in \mathcal{D}_r(N_{i,j})$ such that $\xi_0^{i,j} \in \mathcal{C}_i$, $\xi_{N_{i,j}-1}^{i,j} \in \mathcal{C}_j$, $\xi_n^{i,j} \notin \bigcup_{l \neq i,j} \mathcal{C}_l$ for all $n = 1, \dots, N_{i,j} - 2$ and $\mathcal{S}_{N_{i,j}}(\xi^{i,j}) \leq \mathcal{B}_{i,j} + \varepsilon$.

Define $\delta_{i,j} := \min \left\{ d(\xi_n^{i,j}, \bigcup_{l \neq i,j} \mathcal{C}_l) : n = 1, \dots, N_{i,j} - 2 \right\}$ and $\delta := \min_{i,j \in I} \delta_{i,j}$. By construction, it holds that $\delta > 0$.

Without loss of generality, at the potential cost of reducing ε , we can assume that [Lemma B.1.7](#) can be applied to every $\mathcal{C}_i$, $i = 1, \dots, K$ with $r \leftarrow \varepsilon$ and denote by $\mathcal{W}_i$ the corresponding neighborhoods of $\mathcal{C}_i$. Require that $\mathcal{C}_i^\delta$ be contained in $\mathcal{W}_i \cap \mathcal{U}_{\delta/2}(\mathcal{K}_i)$ for all $i = 1, \dots, K$. Now, given such $\mathcal{C}_i^\delta$ neighborhoods of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, by compactness, there exists $0 < \delta' \leq \delta/2$ such that $\mathcal{U}'_\delta(\mathcal{C}_i)$ is contained in $\mathcal{C}_i^\delta$ for all $i = 1, \dots, N_{\text{crit}}$.

Apply [Lemma B.1.8](#) to $\mathcal{C}_i$, $i = 1, \dots, K$ with $\varepsilon \leftarrow \min(\varepsilon, \delta'/2)$ and denote by N_i the bound on the length of paths obtained.

We are now ready to prove the result. Fix $(i, j) \in \{1, \dots, K\} \times \{1, \dots, N_{\text{crit}}\}$ and consider $x \in \mathcal{C}_i^\delta$. Since $\mathcal{C}_i^\delta \subset \mathcal{W}_i$, there exists $z \in \mathcal{K}_i$ such that $\rho(x, z) < \varepsilon$.

By [Lemma B.1.8](#), there exists $n \leq N$, $\xi \in \mathcal{D}_r(n)$ such that $\xi_0 = z$, $\xi_{N-1} = \xi_0^{i,j}$, $\xi_k \in \mathcal{U}_{\delta'/2}(\mathcal{K}_i)$ for all $k = 1, \dots, n - 2$ and $\mathcal{S}_n(\xi) < \varepsilon$.

Considering the concatenation

$$\zeta := \left(x, \xi_0, \xi_1, \dots, \xi_{n-2}, \xi_{n-1}, \xi_1^{i,j}, \dots, \xi_{N_{i,j}-1}^{i,j} \right) \quad (\text{B.2.25})$$

which is a path of length $n + N_{i,j} + 1$ made of $x \in \mathcal{C}_i^\delta$, then exactly n points in $\mathcal{U}_{\delta'/2}(\mathcal{K}_i)$ then $N_{i,j} - 2$ in $\mathbb{R}^d \setminus \mathcal{U}_{\delta/2}(\mathcal{C}^\delta)$ and $\xi_{N_{i,j}-1}^{i,j} \in \mathcal{K}_j$. Moreover, by construction, $\mathcal{S}_{N+N_{i,j}}(\zeta) \leq \tilde{\mathcal{B}}_{i,j} + 3\varepsilon$. Therefore, if

$$\text{dist}_{n+N_{i,j}+1}(x^\eta, \zeta) < \delta'/2, \quad (\text{B.2.26})$$

with $x_0^\eta = x$, then $x_1^\eta, \dots, x_n^\eta$ are in $\mathcal{U}_{\delta'}(\mathcal{K}_i) \subset \mathcal{C}_i^\delta$ and, since $\delta' \leq \delta/2$, $x_{n+2}^\eta, \dots, x_{n+N_{i,j}-1}^\eta$ are not in $\mathcal{U}_{\delta/4}(\mathcal{C}^\delta)$, and therefore not in $\mathcal{C}^\delta$. Moreover, $x_{n+N_{i,j}}^\eta$ would be in $\mathcal{U}_{\delta'/2}(\mathcal{K}_j) \subset \mathcal{C}_j^\delta$.

Thus, all the paths x^η satisfying (B.2.26) with $x_0^\eta = x$ correspond to exactly one transition of the induced chain $(z_n)_n$ from x to $\mathcal{C}_j^\delta$.

Therefore, using the definition of $\mathbb{Q}_{\mathcal{C}^\delta}$, we have that

$$\begin{aligned} \mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) &\geq \mathbb{P}_x\left(\text{dist}_{n+N_{i,j}+1}(x^\eta, \zeta) < \delta'/2\right) \\ &\geq \exp\left(-\frac{\mathcal{B}_{i,j} + 4\varepsilon}{\eta}\right), \end{aligned} \quad (\text{B.2.27})$$

by [Proposition B.1.3](#). ■

B.2.4 Accelerated induced chain

We will take the following convention, for any $i \in \{1, \dots, N_{\text{crit}}\}$, $\delta > 0$:

$$\mathcal{B}_{i,i}^\delta = \mathcal{B}_{i,i} = \mathcal{B}_{i,i} = 0. \quad (\text{B.2.28})$$

Recall that $\mathbb{Q}_{\mathcal{C}^\delta}$ is the transition probabilities of the induced chain $(z_n)_{n \geq 0}$, defined in [Definition B.12](#). In other words $\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta)$ is the probability that the sequence of (B.1.16) starting from x at time 0 enters $\mathcal{C}_j^\delta$ before any other $\mathcal{C}_l^\delta$ for $l \neq j$. Moreover, K still denotes the number of connected components of the critical set of f that are not part of the global minimum, see [Definition B.5](#) in [Appendix B.1.2](#).

Definition B.13. Given the induced chain $(z_n)_{n \geq 0}$, we define the accelerated induced chain $(z_n^K)_{n \geq 0}$ as follows:

$$z_n^K := z_{nK}. \quad (\text{B.2.29})$$

We denote by $\mathbb{Q}_{\mathcal{C}^\delta}^K$ the transition probabilities of the accelerated induced chain which corresponds to the K -th power of $\mathbb{Q}_{\mathcal{C}^\delta}$: it satisfies¹, for any $x \in \mathcal{C}_i^\delta$, $i \in \{1, \dots, N_{\text{crit}}\}$ and any measurable set $A \subset \mathbb{R}^d$,

$$\mathbb{Q}_{\mathcal{C}^\delta}^K(x, A) = \int \int \cdots \int \mathbb{1}\{z_K \in A\} \mathbb{Q}_{\mathcal{C}^\delta}(x, dz_1) \mathbb{Q}_{\mathcal{C}^\delta}(z_1, dz_2) \cdots \mathbb{Q}_{\mathcal{C}^\delta}(z_{K-1}, dz_K). \quad (\text{B.2.30})$$

We now define transition costs for the accelerated induced chain, which correspond the ones introduced in the main text (B.2.31) in [Appendix B.2](#).

Definition B.14 (Inspired by [Freidlin and Wentzell \(1998b, Chap. 6, §2\)](#)). For $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, $\delta > 0$,

$$\mathcal{B}_{i,j} := \inf \left\{ \mathcal{S}_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 \in \mathcal{C}_i, \xi_{N-1} \in \mathcal{C}_j, \xi_n \notin \bigcup_{l \neq i, j, l > K} \mathcal{C}_l \text{ for all } n = 1, \dots, N-2 \right\} \quad (\text{B.2.31})$$

¹ This equation is sometimes referred to as the Chapman-Kolmogorov equation.

Lemma B.2.5. For any $(i, j) \in \{1, \dots, K\} \times \{1, \dots, N_{\text{crit}}\}$,

$$B_{i,j} = \min \left\{ \sum_{l=0}^{n-2} \mathcal{B}_{i_l, i_{l+1}} : i_0 = i, i_{n-1} = j, i_l \in \{1, \dots, K\} \text{ for } l = 1, \dots, n-2, n \geq 1 \right\} \quad (\text{B.2.32})$$

$$= \min \left\{ \sum_{l=0}^{K-1} \mathcal{B}_{i_l, i_{l+1}} : i_0 = i, i_K = j, i_l \in \{1, \dots, K\} \text{ for } l = 1, \dots, K-1 \right\}. \quad (\text{B.2.33})$$

Proof. We focus on proving the first equality for $B_{i,j}$; the proof for $B_{i,j}$ follows similarly. Let us denote by $\text{RHS}(i, j)$ the right-hand side:

$$\text{RHS}(i, j) := \min \left\{ \sum_{l=0}^{n-2} \mathcal{B}_{i_l, i_{l+1}} : i_0 = i, i_{n-1} = j, i_l \in \{1, \dots, K\} \text{ for } l = 1, \dots, n-2, n \geq 1 \right\} \quad (\text{B.2.34})$$

We will prove that $B_{i,j} = \text{RHS}(i, j)$.

($\geq$): Let us show that $B_{i,j} \geq \text{RHS}(i, j)$. Fix $\varepsilon > 0$. There is a path $\xi \in \mathcal{D}_r(N)$ with $\xi_0 \in \mathcal{C}_i$, $\xi_{N-1} \in \mathcal{C}_j$ that avoids $\bigcup_{l \neq i, j, l > K} \mathcal{C}_l$ so that, by definition of $B_{i,j}$ (Definition B.14),

$$B_{i,j} \geq \mathcal{S}_N(\xi) - \varepsilon. \quad (\text{B.2.35})$$

Let $1 \leq n_1 < n_2 < \dots < n'_N < N$ be the times when ξ belongs to some $\mathcal{C}_l$ with $l \leq K$. Define $i_0 = i$, $i'_N = j$ and for $1 \leq i < N'$, let i_k be such that $\xi_{n_k} \in \mathcal{C}_{i_k}$. Then, by definition of $\mathcal{B}$ (Definition B.11),

$$\mathcal{S}_N(\xi) \geq \sum_{k=0}^{N'-1} \mathcal{B}_{i_k, i_{k+1}}. \quad (\text{B.2.36})$$

Combining (B.2.35) and (B.2.36) and taking yields

$$B_{i,j} \geq \sum_{k=0}^{N'-1} \mathcal{B}_{i_k, i_{k+1}} - \varepsilon \quad (\text{B.2.37})$$

$$\geq \text{RHS}(i, j) - \varepsilon, \quad (\text{B.2.38})$$

since this sequence $(i_k)_{k=0}^{N'}$ satisfies $i_0 = i$, $i'_N = j$ and $i_k \in \{1, \dots, K\}$ for $1 \leq k < N'$.

($\leq$): Let us show that $B_{i,j} \leq \text{RHS}(i, j)$. Take $\varepsilon > 0$ and let $(i_k)_{k=0}^{N'}$ be a sequence achieving the minimum in $\text{RHS}(i, j)$ up to ε , i.e.,

$$\sum_{k=0}^{N'-1} \mathcal{B}_{i_k, i_{k+1}} \leq \text{RHS}(i, j) + \varepsilon. \quad (\text{B.2.39})$$

Take $\delta > 0$ small enough such that $\mathcal{U}_\delta(\mathcal{C}_l)$, $l = 1, \dots, N_{\text{crit}}$ are pairwise disjoint.

For each $0 \leq k < N'$, by definition of $\mathcal{B}$ (Definition B.11), there exists $\delta_k \in (0, \delta)$ and a path $\xi^k \in \mathcal{D}_r(N_k)$ such that: (i) $\xi_0^k \in \mathcal{C}_{i_k}$, (ii) $\xi_{N_k-1}^k \in \mathcal{C}_{i_{k+1}}$, (iii) $\xi_n^k \notin \bigcup_{l \neq i_k, i_{k+1}} \mathcal{C}_l$ for all $n = 1, \dots, N_k - 2$, and (iv) $\mathcal{S}_{N_k}(\xi^k) \leq \mathcal{B}_{i_k, i_{k+1}} + \varepsilon/N'$.

By Lemma B.1.8, for each $0 \leq k < N' - 1$, there exists $\tilde{\xi}^k \in \mathcal{D}_r(\tilde{N}_k)$ such that: (i) $\tilde{\xi}_0^k = \xi_{N_k-1}^k$, (ii) $\tilde{\xi}_{\tilde{N}_k-1}^k = \xi_0^{k+1}$, (iii) $\tilde{\xi}_n^k \in \mathcal{U}_{\delta/2}(\mathcal{C}_{i_{k+1}})$ and therefore $\tilde{\xi}_n^k \notin \bigcup_{l \neq i_{k+1}} \mathcal{C}_l$ for all $n = 1, \dots, \tilde{N}_k - 2$, and (iv) $\mathcal{S}_{\tilde{N}_k}(\tilde{\xi}^k) \leq \varepsilon/N'$.

Concatenating the paths $\xi^0, \tilde{\xi}^0, \xi^1, \tilde{\xi}^1, \dots, \xi^{N'-1}$ yields a path ζ that: (i) starts in $\mathcal{C}_i$, (ii) ends in $\mathcal{C}_j$, (iii) avoids $\bigcup_{l \neq i, j, l > K} \mathcal{C}_l$, and (iv) has total cost at most $\sum_{k=0}^{N'-1} \mathcal{B}_{i_k, i_{k+1}} + 3\varepsilon$. Therefore, by definition of $B_{i,j}$ (Definition B.14),

$$\begin{aligned} B_{i,j} &\leq \sum_{k=0}^{N'-1} \mathcal{B}_{i_k, i_{k+1}} + 3\varepsilon \\ &\leq \text{RHS}(i, j) + 4\varepsilon. \end{aligned} \quad (\text{B.2.40})$$

The second equality in each case follows from the fact that optimal paths between different components can be chosen without cycles (since all costs are non-negative), and therefore their length can be bounded by the number of components K . ■

Lemma B.2.6 (Upper bound on accelerated induced chain transition probability). *For any $\varepsilon > 0$, $A > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such that for all $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, $x \in \mathcal{C}_i^\delta$, $0 < \eta < \eta_0$,*

$$\mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) \leq \begin{cases} \exp\left(-\frac{B_{i,j}}{\eta} + \frac{\varepsilon}{\eta}\right) & \text{if } B_{i,j} < +\infty, \\ \exp\left(-\frac{A}{\eta}\right) & \text{otherwise.} \end{cases} \quad (\text{B.2.41})$$

Proof. Fix $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$. By Lemma B.2.5, we have that

$$B_{i,j} = \min \left\{ \sum_{l=0}^{K-1} \mathcal{B}_{i_l, i_{l+1}} : i_0 = i, i_K = j, i_l \in \{1, \dots, K\} \text{ for } l = 1, \dots, K-1 \right\}. \quad (\text{B.2.42})$$

First, let us consider the case where $B_{i,j} < +\infty$. Take $\varepsilon > 0$ and let $(i_k)_{k=0}^K$ be a sequence achieving the minimum in (B.2.42) up to ε , i.e.,

$$\sum_{k=0}^{K-1} \mathcal{B}_{i_k, i_{k+1}} \leq B_{i,j} + \varepsilon. \quad (\text{B.2.43})$$

By definition of the accelerated induced chain (Definition B.13), we have that for any $x \in \mathcal{C}_i^\delta$,

$$\mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) = \sum_{i_1, \dots, i_{K-1} \in \{1, \dots, K\}} \int_{\mathcal{C}_{i_1}^\delta} \cdots \int_{\mathcal{C}_{i_{K-1}}^\delta} \prod_{k=0}^{K-1} \mathbb{Q}_{\mathcal{C}^\delta}(x_k, dx_{k+1}) \mathbb{Q}_{\mathcal{C}^\delta}(x_K, \mathcal{C}_j^\delta), \quad (\text{B.2.44})$$

where we define $x_0 = x$.

Therefore, combining (B.2.44) with Lemma B.2.3 and (B.2.43), we obtain that for any $x \in \mathcal{C}_i^\delta$,

$$\begin{aligned} \mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) &\leq \sum_{i_1, \dots, i_{K-1} \in \{1, \dots, K\}} \prod_{k=0}^{K-1} \exp\left(-\frac{\mathcal{B}_{i_k, i_{k+1}}}{\eta} + \frac{\varepsilon}{\eta K}\right) \\ &\leq K^{K-1} \exp\left(-\frac{B_{i,j}}{\eta} + \frac{2\varepsilon}{\eta}\right). \end{aligned} \quad (\text{B.2.45})$$

For the case where $B_{i,j} = +\infty$, by the same reasoning but using the alternative bound from [Lemma B.2.3](#), we get

$$\mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) \leq K^{K-1} \exp\left(-\frac{A}{\eta}\right). \quad (\text{B.2.46})$$

Taking η_0 small enough so that $K^{K-1} \leq \exp(\varepsilon/\eta)$ for all $\eta \leq \eta_0$, the result follows from [\(B.2.45\)](#) and [\(B.2.46\)](#). ■

Lemma B.2.7 (Lower bound on accelerated induced chain transition probability). *For any $\varepsilon > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such that for all $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, $x \in \mathcal{C}_i^\delta$, $0 < \eta < \eta_0$,*

$$\mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) \geq \exp\left(-\frac{B_{i,j}}{\eta} - \frac{\varepsilon}{\eta}\right). \quad (\text{B.2.47})$$

The proof of [Lemma B.2.7](#) is very similar to the proof of [Lemma B.2.6](#) and is therefore omitted.

We end this section by restating a result from ([Azizian et al., 2024b](#)) that provides sufficient conditions for $B_{i,j}$ to be finite for any $(i, j) \in \{1, \dots, K\} \times \{1, \dots, N_{\text{crit}}\}$.

Lemma B.2.8 (([Azizian et al., 2024b](#), Lem. D.12)). *Consider $x, x' \in \mathbb{R}^d$ and assume that there exists $T \in \mathbb{N}$, $\gamma \in \mathcal{C}^1([0, T], \mathbb{R}^d)$ such that $\gamma_0 = x$, $\gamma_T = x'$, $\gamma_n \notin \bigcup_{l \neq i, j, l > K} \mathcal{C}_l$ for all $n = 1, \dots, T-1$ and, for every $t \in [0, T]$, $\nabla f(\gamma_t)$ is in the interior of the closed convex hull of the support of $U(\gamma_t, \omega)$, i.e.,*

$$\nabla f(\gamma_t) \in \text{int} \overline{\text{conv}} \text{supp } U(\gamma_t, \omega). \quad (\text{B.2.48})$$

Then, $\mathcal{B}(x, x') < +\infty$.

B.2.5 Initial transition

Definition B.15 (Inspired by [Freidlin and Wentzell \(1998b, Chap. 6, §2\)](#)). For $j \in \{1, \dots, N_{\text{crit}}\}$, $\delta > 0$, $x \in \mathbb{R}^d$,

$$\mathcal{B}_{x,j} := \inf \left\{ \mathcal{S}_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 = x, \xi_{N-1} \in \mathcal{C}_j, \xi_n \notin \bigcup_{l \neq i, j} \mathcal{C}_l \text{ for all } n = 1, \dots, N-2 \right\} \quad (\text{B.2.49})$$

$$\mathcal{B}_{x,j}^\delta := \inf \left\{ \mathcal{S}_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 = x, \xi_{N-1} \in \mathcal{U}_\delta(\mathcal{C}_j), \xi_n \notin \bigcup_{l \neq i, j} \mathcal{C}_l \text{ for all } n = 1, \dots, N-2 \right\}. \quad (\text{B.2.50})$$

The proof of the following lemmas are very similar to the ones of [Lemma B.2.2](#), [Lemma B.2.3](#), and [Lemma B.2.4](#) and are therefore omitted.

Lemma B.2.9. *For any $j \in \{1, \dots, N_{\text{crit}}\}$, $\delta > 0$, $x \in \mathbb{R}^d$,*

$$\mathcal{B}_{x,j} = \lim_{\delta \rightarrow 0} \mathcal{B}_{x,j}^\delta. \quad (\text{B.2.51})$$

Lemma B.2.10. For any $\varepsilon > 0$, $x \in \mathbb{R}^d$, for any neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$ small enough, there exists $\eta_0 > 0$ such that for all $j \in \{1, \dots, N_{\text{crit}}\}$, $0 < \eta < \eta_0$,

$$\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) \geq \exp\left(-\frac{\mathcal{B}_{x,j}}{\eta} - \frac{\varepsilon}{\eta}\right). \quad (\text{B.2.52})$$

Lemma B.2.11. For any $\varepsilon > 0$, $A > 0$, $x \in \mathbb{R}^d$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such that for all $j \in \{1, \dots, N_{\text{crit}}\}$, $0 < \eta < \eta_0$,

$$\mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) \leq \begin{cases} \exp\left(-\frac{\mathcal{B}_{x,j}}{\eta} + \frac{\varepsilon}{\eta}\right) & \text{if } \mathcal{B}_{x,j} < +\infty, \\ \exp\left(-\frac{A}{\eta}\right) & \text{otherwise.} \end{cases} \quad (\text{B.2.53})$$

Following the same methodology as for transitions between critical components, we now establish similar results for transitions starting from an arbitrary initial state.

Definition B.16 (Inspired by [Freidlin and Wentzell \(1998b, Chap. 6, §2\)](#)). For $x \in \mathbb{R}^d$, $j \in \{1, \dots, N_{\text{crit}}\}$, $\delta > 0$,

$$B_{x,j} := \inf \left\{ S_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 = x, \xi_{N-1} \in \mathcal{C}_j, \xi_n \notin \bigcup_{l \neq j, l > K} \mathcal{C}_l \text{ for all } n = 1, \dots, N-2 \right\} \quad (\text{B.2.54})$$

This definition mirrors the one of [Definition B.14](#), adapting it to account for an arbitrary initial state rather than starting from a critical component. We can then establish the following decomposition result, analogous to [Lemma B.2.5](#).

Lemma B.2.12. For any $x \in \mathbb{R}^d$, $j \in \{1, \dots, N_{\text{crit}}\}$,

$$B_{x,j} = \min \left\{ B_{x,i_1} + \sum_{l=1}^{n-2} \mathcal{B}_{i_l, i_{l+1}} : i_{n-1} = j, i_l \in \{1, \dots, K\} \text{ for } l = 1, \dots, n-2, n \geq 1 \right\} \quad (\text{B.2.55})$$

$$= \min \left\{ B_{x,i_1} + \sum_{l=1}^{K-1} \mathcal{B}_{i_l, i_{l+1}} : i_K = j, i_l \in \{1, \dots, K\} \text{ for } l = 1, \dots, K-1 \right\} \quad (\text{B.2.56})$$

The proof follows the same arguments as in [Lemma B.2.5](#), replacing the initial component with the given initial state.

These structural results allow us to establish bounds on the transition probabilities from an arbitrary initial state, paralleling those of [Lemma B.2.6](#) and [Lemma B.2.7](#).

Lemma B.2.13 (Upper bound on accelerated induced chain initial transition probability). For any $x \in \mathbb{R}^d$, for any $\varepsilon > 0$, $A > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such that for all $j \in \{1, \dots, N_{\text{crit}}\}$, $0 < \eta < \eta_0$,

$$\mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) \leq \begin{cases} \exp\left(-\frac{\mathcal{B}_{x,j}}{\eta} + \frac{\varepsilon}{\eta}\right) & \text{if } \mathcal{B}_{x,j} < +\infty, \\ \exp\left(-\frac{A}{\eta}\right) & \text{otherwise.} \end{cases} \quad (\text{B.2.57})$$

Lemma B.2.14 (Lower bound on accelerated induced chain initial transition probability). *For any $x \in \mathbb{R}^d$, for any $\varepsilon > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such that for all $j \in \{1, \dots, N_{\text{crit}}\}$, $0 < \eta < \eta_0$,*

$$\mathbf{Q}^K_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) \geq \exp\left(-\frac{B_{x,j}}{\eta} - \frac{\varepsilon}{\eta}\right). \quad (\text{B.2.58})$$

The proofs of these last two lemmas follow very closely those of [Lemma B.2.6](#) and [Lemma B.2.7](#), with the main modification being the treatment of the initial state instead of an initial component. The key arguments involving the Chapman-Kolmogorov equation and the handling of the intermediate transitions remain essentially unchanged.

B.2.6 Transition time

Let us begin with a preliminary escape result.

Lemma B.2.15. *Given C a connected components of $\mathcal{K}_{\text{crit}}$, for any $\varepsilon > 0$, for $\eta > 0$ small enough and $\mathcal{C}^\delta$ a small enough neighborhood of C , it holds that, for any $x \in \mathcal{C}^\delta$,*

$$\mathbb{E}_x \left[\sigma_{\mathcal{C}^\delta}^\eta \right] \leq \exp\left(\frac{\varepsilon}{\eta}\right). \quad (\text{B.2.59})$$

Proof. Fix $\varepsilon > 1$. By [Lemma B.1.7](#), $\mathcal{W}_\varepsilon(C)$ is an open neighborhood of C . Since C is closed, $\mathcal{W}_\varepsilon(C) \setminus C$ is not empty and open. Therefore, there exists $x_3 \in \mathcal{W}_\varepsilon(C) \setminus C$ and $\delta > 0$ such that $\mathbb{B}(x_3, \delta) \subset \mathcal{W}_\varepsilon(C) \setminus C$. By construction, there exists $x_2 \in C$ such that $\rho(x_2, x_3) < \varepsilon$.

Let us now apply [Lemma B.1.8](#) to obtain that there exists $N \geq 2$ such that, for any $x_1 \in C$, there exists $n \leq N$, $\xi \in \mathcal{D}_r(N)$ such that $\xi_0 = x_1$, $\xi_{N-1} = x_2$, $\mathcal{S}_n(\xi) < \varepsilon$.

Require then that $\mathcal{C}^\delta$ be contained in $\mathcal{W}_\varepsilon(C) \setminus \overline{\mathbb{B}}(x_3, \delta/2)$ which is an open neighborhood of C .

Take $x \in \mathcal{C}^\delta$. Since $\mathcal{C}^\delta$ is in particular contained in $\mathcal{W}_\varepsilon(C)$, there exists $x_1 \in C$ such that $\rho(x, x_1) < \varepsilon$. By our application of [Lemma B.1.8](#) above, there exists $n \leq N$, $\xi \in \mathcal{D}_r(n)$ such that $\xi_0 = x_1$, $\xi_{n-1} = x_2$, $\mathcal{S}_n(\xi) < \varepsilon$. Now consider the path $\zeta \in \mathcal{D}_r(n+2)$ defined by

$$\zeta = (x, \xi_0, \xi_1, \dots, \xi_{n-1}, x_3), \quad (\text{B.2.60})$$

which satisfies

$$\mathcal{S}_{n+3}(\zeta) = \rho(x, x_1) + \mathcal{S}_n(\xi) + \rho(x_2, x_3) < 3\varepsilon. \quad (\text{B.2.61})$$

If a trajectory of SGD x^η with $x^\eta_0 = x$ satisfies $\text{dist}_{n+2}(x^\eta, \zeta) < \delta/2$, then x^η_{n+1} is inside $\mathbb{B}(x_3, \delta/2)$ and therefore not in $\mathcal{C}^\delta$.

Hence, we have, for $\eta > 0$ small enough,

$$\begin{aligned} \mathbf{P}_x(\sigma_{\mathcal{C}^\delta}^\eta < N+3) &\geq \mathbf{P}_x(\sigma_{\mathcal{C}^\delta}^\eta < n+3) \\ &\geq \mathbf{P}_x(\text{dist}_{n+2}(x^\eta, \zeta) < \delta/2) \\ &\geq \exp\left(-\frac{\mathcal{S}_{n+2}(\zeta) + \varepsilon}{\eta}\right) \\ &\geq \exp\left(-\frac{4\varepsilon}{\eta}\right), \end{aligned} \quad (\text{B.2.62})$$

where we invoked [Proposition B.1.3](#) and used [\(B.2.61\)](#).

For any $x \in \mathcal{C}^\delta$, for $n \geq 2$, the (weak) Markov property yields that

$$\begin{aligned} \mathbb{P}_x(\sigma_{\mathcal{C}^\delta}^\eta > n(N+2)) &= \mathbb{E}_x \left[\mathbb{1}\{\sigma_{\mathcal{C}^\delta}^\eta > N+2\} \mathbb{E}_{x_{N+2}}^\eta \left[\mathbb{1}\{\sigma_{\mathcal{C}^\delta}^\eta > N+2\} \dots \mathbb{E}_{x_{(n-2)(N+1)}}^\eta \left[\mathbb{1}\{\sigma_{\mathcal{C}^\delta}^\eta > N+2\} \right] \dots \right] \right] \\ &\leq \left(2 - \exp\left(-\frac{4\varepsilon}{\eta}\right) \right)^n, \end{aligned} \quad (\text{B.2.63})$$

with [\(B.2.62\)](#).

We can now finally estimate the expectation of $\sigma_{\mathcal{C}^\delta}^\eta$: for $x \in \mathcal{C}^\delta$,

$$\begin{aligned} \mathbb{E}_x[\sigma_{\mathcal{C}^\delta}^\eta] &= \sum_{n=1}^{\infty} \mathbb{P}_x(\sigma_{\mathcal{C}^\delta}^\eta > n) \\ &= \sum_{n=1}^{\infty} \sum_{k=0}^N \mathbb{P}_x(\sigma_{\mathcal{C}^\delta}^\eta > n(N+1) + k) \\ &\leq (N+2) \sum_{n=0}^{\infty} \mathbb{P}_x(\sigma_{\mathcal{C}^\delta}^\eta > n(N+1)) \\ &\leq (N+2) \sum_{n=0}^{\infty} \left(1 - \exp\left(-\frac{4\varepsilon}{\eta}\right) \right)^n \\ &= (N+2) \exp\left(\frac{4\varepsilon}{\eta}\right), \end{aligned} \quad (\text{B.2.64})$$

where we used [\(B.2.63\)](#).

Finally, taking $\eta > 0$ small enough so that $N+1 \leq \exp\left(\frac{\varepsilon}{\eta}\right)$ yields that

$$\mathbb{E}_x[\sigma_{\mathcal{C}^\delta}^\eta] \leq \exp\left(\frac{5\varepsilon}{\eta}\right), \quad (\text{B.2.65})$$

which concludes the proof. $\blacksquare$

Lemma B.2.16. *For any $\mathcal{X}_0 \subset \mathbb{R}^d$ compact, for any $\varepsilon > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such for all $0 < \eta < \eta_0$, for any $x \in \mathcal{X}_0 \setminus \bigcup_{i=K+1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$,*

$$1 \leq \mathbb{E}_x[\tau_1^\eta] \leq e^{\frac{\varepsilon}{\eta}}. \quad (\text{B.2.66})$$

Proof. Let us begin with the LHS inequality. There are two cases: either x belongs to $\bigcup_{i=1}^K \mathcal{C}_i^\delta$ or not. If x belongs to $\bigcup_{i=1}^K \mathcal{C}_i^\delta$, then by definition $\sigma_1^\eta \geq 1$ and therefore $\tau_1^\eta \geq 1$. If x does not belong to $\bigcup_{i=1}^K \mathcal{C}_i^\delta$, then $\sigma_1^\eta = 0$ but necessarily $\tau_1^\eta \geq 1$. Hence, in all cases, $\tau_1^\eta \geq 1$ so that the LHS inequality holds.

We now turn to the RHS inequality. As before, we will separate the proof into two cases: either x belongs to $\bigcup_{i=1}^K \mathcal{C}_i^\delta$ or not. If x belongs to some $\mathcal{C}_i^\delta$ for some $i \in \{1, \dots, K\}$ then, by [Lemma B.2.15](#), we have that

$$\mathbb{E}_x[\sigma_1^\eta] = \mathbb{E}_x[\sigma_{\mathcal{C}_i^\delta}^\eta] \leq e^{\frac{\varepsilon}{\eta}}. \quad (\text{B.2.67})$$

In particular, σ_1^η is finite almost surely. Therefore, the strong Markov property implies that

$$\mathbb{E}_x[\tau_1^\eta] = \mathbb{E}_x[\sigma_1^\eta] + \mathbb{E}_x[\mathbb{E}_{x_{\sigma_1^\eta}}^\eta[\tau_{\mathcal{C}^\delta}^\eta]]. \quad (\text{B.2.68})$$

Applying [Lemma B.1.4](#) with $\mathcal{U} \leftarrow \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$ and $\mathbb{R}^d \leftarrow \mathcal{X}_0$, and using Jensen's inequality, we obtain that

$$\mathbb{E}_{x_{\sigma_1}^\eta}[\tau_{\mathcal{C}^\delta}^\eta] \leq a + \frac{b\eta_0}{\alpha_0} =: c. \quad (\text{B.2.69})$$

Plugging this bound into [\(B.2.68\)](#) and using [\(B.2.67\)](#) yields

$$\mathbb{E}_x[\tau_1^\eta] \leq e^{\frac{\varepsilon}{\eta}} + c, \quad (\text{B.2.70})$$

which concludes the proof of this case.

Finally, if x does not belong to $\bigcup_{i=1}^K \mathcal{C}_i^\delta$, then one only needs to apply [Lemma B.1.4](#) as above to obtain the result. $\blacksquare$

Corollary B.2.17. *For any $\mathcal{X}_0 \subset \mathbb{R}^d$ compact, for any $N \geq 1$, for any $\varepsilon > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$, for any $x \in \mathcal{X}_0 \setminus \bigcup_{i=K+1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$,*

$$1 \leq \mathbb{E}_x[\tau_N^\eta] \leq e^{\frac{\varepsilon}{\eta}}. \quad (\text{B.2.71})$$

Proof. The lower bound follows directly from the definition of τ_N^η since $\tau_k^\eta \geq k$ for all $k \geq 1$ by construction.

For the upper bound, first note that we can write τ_N^η as a telescoping sum:

$$\tau_N^\eta = \sum_{k=0}^{N-1} (\tau_{k+1}^\eta - \tau_k^\eta), \quad (\text{B.2.72})$$

with the convention that $\tau_0^\eta = 0$. Therefore,

$$\mathbb{E}_x[\tau_N^\eta] = \sum_{k=0}^{N-1} \mathbb{E}_x[\tau_{k+1}^\eta - \tau_k^\eta]. \quad (\text{B.2.73})$$

By the strong Markov property applied at time τ_k^η for each term, we have

$$\begin{aligned} \mathbb{E}_x[\tau_{k+1}^\eta - \tau_k^\eta] &= \mathbb{E}_x[\mathbb{E}_{x_{\tau_k}^\eta}[\tau_1^\eta]] \\ &\leq e^{\frac{\varepsilon}{\eta}}, \end{aligned} \quad (\text{B.2.74})$$

where the inequality follows from [Lemma B.2.16](#) since $x_{\tau_k}^\eta$ belongs to $\bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$ by definition of τ_k^η for all $k \geq 1$ (and for $k = 0$, we can apply [Lemma B.2.16](#) directly to the initial point x).

Summing over k from 0 to $N - 1$ yields

$$\begin{aligned} \mathbb{E}_x[\tau_N^\eta] &\leq N e^{\frac{\varepsilon}{\eta}} \\ &\leq e^{\frac{2\varepsilon}{\eta}}, \end{aligned} \quad (\text{B.2.75})$$

for η small enough. $\blacksquare$

B.3 FINITE-TIME ANALYSIS

B.3.1 Markov Chains on finite state spaces

In this subsection, we introduce the necessary notation and then restate a key lemma from [Freidlin and Wentzell \(1998b, Chap. 6,§3\)](#).

Consider a finite set $\mathcal{V}$ and $\mathcal{Q} \subseteq \mathcal{V}$. Denote by $r := \text{card } \mathcal{V} \setminus \mathcal{Q}$.

Given $p : (\mathcal{V} \setminus \mathcal{Q}) \times \mathcal{V} \rightarrow [0, 1]$, we define the probability of a g with edges in $(\mathcal{V} \setminus \mathcal{Q}) \times \mathcal{V}$ as

$$\pi(g, p) := \prod_{i \rightarrow j \in g} p(i \rightarrow j). \quad (\text{B.3.1})$$

A graph g consisting of arrows $j \rightarrow k$ with $j \in \mathcal{V} \setminus \mathcal{Q}$, $k \in \mathcal{V}$, $j \neq k$ is called a $\mathcal{Q}$ -graph on $\mathcal{V}$ if

- (i) Every vertex in $\mathcal{V} \setminus \mathcal{Q}$ has exactly one outgoing arrow.
- (ii) There are no cycles, or, equivalently, from every vertex in $\mathcal{V} \setminus \mathcal{Q}$ there is a directed path to a vertex in $\mathcal{Q}$.

The set of $\mathcal{Q}$ -graphs is denoted by $\mathcal{G}(\mathcal{Q})$.

We denote by $\mathcal{G}(i \nearrow \mathcal{Q})$ the set of graphs with exactly $\text{card}(\mathcal{V} \setminus \mathcal{Q}) - 1$ edges from $\mathcal{V} \setminus \mathcal{Q}$ to $\mathcal{V}$, no cycles and no path from i to $\mathcal{Q}$. Note that, equivalently, this set is made of all $\mathcal{Q}$ -graphs from which a single edge from the path from i to $\mathcal{Q}$ has been removed.

Lemma B.3.1 ([Freidlin and Wentzell \(1998b, Chap. 6,§3,Lem. 3.4\)](#)). *Consider a Markov Chain on state space $X = \bigcup_{i \in \mathcal{V}} X_i$ with X_i , $i \in \mathcal{V}$ disjoint and non-empty, with transition probabilities that satisfy: for $i \in \mathcal{V} \setminus \mathcal{Q}$, $j \in \mathcal{V}$ with $i \neq j$,*

$$\forall x \in X_i, \quad a^{-1}p_{ij} \leq \mathbf{P}(x, X_j) \leq ap_{ij}, \quad (\text{B.3.2})$$

for some $a \geq 1$, $p_{ij} > 0$. For $i \in \mathcal{V} \setminus \mathcal{Q}$, the expected time to reach $\mathcal{Q}$ when starting at x denoted by $m_{\mathcal{Q}}(x)$ satisfies: for all $x \in X_i$,

$$a^{3 \times 4^r} \frac{\sum_{g \in \mathcal{G}(i \nearrow \mathcal{Q})} \pi(g, \bar{p})}{\sum_{g \in \mathcal{G}} \pi(g, \bar{p})} \leq m_{\mathcal{Q}}(x) \leq a^{3 \times 4^r} \frac{\sum_{g \in \mathcal{G}(i \nearrow \mathcal{Q})} \pi(g, \underline{p})}{\sum_{g \in \mathcal{G}} \pi(g, \underline{p})} \quad (\text{B.3.3})$$

B.3.2 Hitting time of the accelerated process

We will instantiate the lemmas of the previous section [Appendix B.3.1](#) with $\mathcal{V} = \{1, \dots, N_{\text{crit}}\}$, $\mathcal{Q} = \{K+1, \dots, N_{\text{crit}}\}$ and $X_i = \mathcal{C}_i^\delta$ for $i \in \{1, \dots, N_{\text{crit}}\}$. We have $r = N_{\text{crit}} - K$. Note that the notation $\mathcal{Q}$ is in line with the notation of the main text where we denote by $\mathcal{Q}$ both the set of indices $\{K+1, \dots, N_{\text{crit}}\}$ and the set of union of the corresponding components $\bigcup_{j \in \mathcal{Q}} \mathcal{C}_j = \bigcup_{j=1}^{N_{\text{targ}}} \mathcal{Q}_j = \arg \min f$.

Define by $\tau_{\mathcal{Q}}^\eta := \tau_{\bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta}^\eta$ the hitting time of $\bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$ for the accelerated SGD process. We also consider the induced chain $(z_n)_{n \geq 0}$ on $\bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$ defined in [Definition B.12](#) as well as its accelerated version defined in [Definition B.13](#) and denote by $\tilde{\tau}_{\mathcal{Q}}^K$ the hitting time of $\bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$ for this accelerated induced chain. These two hitting times are related by the following lemma, which is a key consequence of [Lemma B.2.16](#).

Lemma B.3.2. For any $\varepsilon > 0$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such for all $0 < \eta < \eta_0$, for any $x \in \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$,

$$\mathbb{E}_x[\tilde{\tau}_Q^K] \leq \mathbb{E}_x[\tau_Q^\eta] \leq \mathbb{E}_x[\tilde{\tau}_Q^K] \times e^{\frac{\varepsilon}{\eta}}. \quad (\text{B.3.4})$$

Proof. First, if x belongs to $\bigcup_{i=K+1}^{N_{\text{crit}}} \mathcal{C}_i^\delta = \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$, then $\tau_Q^\eta = 0$ and $\tilde{\tau}_Q^K = 0$ so the statement holds trivially.

We now consider the case where x belongs to $\bigcup_{i=1}^K \mathcal{C}_i^\delta$. Since $\mathcal{C}_1, \dots, \mathcal{C}_K$ are compact, we can require that $\bigcup_{i=1}^K \mathcal{C}_i^\delta$ be relatively compact. We can now apply [Corollary B.2.17](#) to the accelerated process to obtain that with $\mathcal{X}_0 \leftarrow \text{cl} \bigcup_{i=1}^K \mathcal{C}_i^\delta$, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such for all $0 < \eta < \eta_0$, for any $x \in \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$, we have,

$$1 \leq \mathbb{E}_x[\tau_K^\eta] \leq e^{\frac{\varepsilon}{\eta}}. \quad (\text{B.3.5})$$

We now have that, by [Definition B.12](#),

$$\begin{aligned} \mathbb{E}_x[\tau_Q^\eta] &= \mathbb{E}_x \left[\sum_{n=0}^{\infty} \mathbb{1}_{\{x_{\tau_{nK}^\eta}^\eta \notin \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta\}} (\tau_{(n+1)K}^\eta - \tau_{nK}^\eta) \right] \\ &= \sum_{n=0}^{\infty} \mathbb{E}_x \left[\mathbb{1}_{\{x_{\tau_{nK}^\eta}^\eta \notin \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta\}} (\tau_{(n+1)K}^\eta - \tau_{nK}^\eta) \right] \\ &= \sum_{n=0}^{\infty} \mathbb{E}_x \left[\mathbb{1}_{\{x_{\tau_{nK}^\eta}^\eta \notin \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta\}} \mathbb{E}_{x_{\tau_{nK}^\eta}^\eta}[\tau_K^\eta] \right], \end{aligned} \quad (\text{B.3.6})$$

where we used the strong Markov property in the last equality, since τ_1^η is always finite almost surely by [\(B.3.5\)](#).

Combining [\(B.3.6\)](#) with [\(B.3.5\)](#), we obtain the bound:

$$\mathbb{E}_x[\tau_Q^\eta] = \sum_{n=0}^{\infty} \mathbb{E}_x \left[\mathbb{1}_{\{x_{\tau_n^\eta}^\eta \notin \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta\}} \right] \leq \mathbb{E}_x[\tau_Q^\eta] \leq \sum_{n=0}^{\infty} \mathbb{E}_x \left[\mathbb{1}_{\{x_{\tau_n^\eta}^\eta \notin \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta\}} \right] \times e^{\frac{\varepsilon}{\eta}} = \mathbb{E}_x[\tilde{\tau}_Q^K] \times e^{\frac{\varepsilon}{\eta}}, \quad (\text{B.3.7})$$

which yields the result. $\blacksquare$

Assumption B.17. For any $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$,

$$B_{i,j} < +\infty. \quad (\text{B.3.8})$$

From now on, we will assume that [Assumption B.17](#) holds.

Definition B.18. For $i \in \mathcal{V} \setminus \mathcal{Q}$, we define the following quantities:

$$E(i \nearrow \mathcal{Q}) := \min \left\{ \sum_{k \rightarrow l \in \mathcal{G}} B_{k,l} : g \in \mathcal{G}(i \nearrow \mathcal{Q}) \right\} \quad (\text{B.3.9})$$

$$E(\mathcal{Q}) := \min \left\{ \sum_{k \rightarrow l \in \mathcal{G}} B_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\} \quad (\text{B.3.10})$$

$$E(\mathcal{Q} | i) := E(\mathcal{Q}) - E(i \nearrow \mathcal{Q}). \quad (\text{B.3.11})$$

Note that [Assumption B.17](#) ensures that all these quantities are finite.

Lemma B.3.3. *For any $i \in \mathcal{V} \setminus \mathcal{Q}$, for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$, for any $x \in \mathcal{C}_i^\delta$,*

$$e^{\frac{E(\mathcal{Q}|i)-\varepsilon}{\eta}} \leq \mathbb{E}_x[\tilde{\tau}_{\mathcal{Q}}^K] \leq e^{\frac{E(\mathcal{Q}|i)+\varepsilon}{\eta}}. \quad (\text{B.3.12})$$

Proof. Fix $i \in \mathcal{V} \setminus \mathcal{Q}$. We will apply [Lemma B.3.1](#) to the accelerated induced chain.

Let us first verify the assumptions of these lemmas. By [Lemma B.2.6](#) and [Lemma B.2.7](#), for any $\varepsilon > 0$, for small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there exists $\eta_0 > 0$ such that for all $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, $x \in \mathcal{C}_i^\delta$, $0 < \eta < \eta_0$:

$$\mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) \leq \exp\left(-\frac{B_{i,j}}{\eta} + \frac{\varepsilon}{\eta}\right), \quad (\text{B.3.13})$$

$$\mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) \geq \exp\left(-\frac{B_{i,j}}{\eta} - \frac{\varepsilon}{\eta}\right). \quad (\text{B.3.14})$$

Note that [Assumption B.17](#) ensures that both $B_{i,j}$ and $B_{j,i}$ are finite.

We define, for any $i \in \mathcal{V} \setminus \mathcal{Q}$, $j \in \mathcal{V}$ with $i \neq j$:

$$p_{ij} := \exp\left(-\frac{B_{i,j}}{\eta}\right) \quad \text{and} \quad p_{ji} := \exp\left(-\frac{B_{j,i}}{\eta}\right). \quad (\text{B.3.15})$$

Let us verify the conditions of [Lemma B.3.1](#) with $a := e^{\varepsilon/\eta}$:

By [\(B.3.13\)](#) and [\(B.3.14\)](#), for all $i \in \mathcal{V} \setminus \mathcal{Q}$, $j \in \mathcal{V}$ with $i \neq j$, $x \in \mathcal{C}_i^\delta$:

$$ap_{ij} \leq \mathbb{Q}_{\mathcal{C}^\delta}^K(x, \mathcal{C}_j^\delta) \leq ap_{ji}. \quad (\text{B.3.16})$$

Now we can apply [Lemma B.3.1](#) to obtain that for $x \in \mathcal{C}_i^\delta$:

$$a^{-3 \times 4^r} e^{\frac{E(\mathcal{Q}|i)}{\eta}} \leq \mathbb{E}_x[\tilde{\tau}_{\mathcal{Q}}^K] \leq a^{3 \times 4^r} e^{\frac{E(\mathcal{Q}|i)}{\eta}}. \quad (\text{B.3.17})$$

Recalling that $a = e^{\varepsilon/\eta}$, the bounds in [\(B.3.17\)](#) become:

$$e^{\frac{E(\mathcal{Q}|i)-3 \times 4^r \varepsilon}{\eta}} \leq \mathbb{E}_x[\tilde{\tau}_{\mathcal{Q}}^K] \leq e^{\frac{E(\mathcal{Q}|i)+3 \times 4^r \varepsilon}{\eta}}, \quad (\text{B.3.18})$$

which concludes the proof since $\varepsilon > 0$ was arbitrary. ■

Definition B.19. Let us now define, for $x \in \mathbb{R}^d$,

$$E(\mathcal{Q}|x) = \left[\max_{i \in \mathcal{V} \setminus \mathcal{Q}} (E(\mathcal{Q}|i) - B_{xi}) \right]_+. \quad (\text{B.3.19})$$

with the convention that these quantities are zero if $x \in \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$ or equal to the quantities defined in [Definition B.18](#) if $x \in \bigcup_{i=1}^K \mathcal{C}_i^\delta$.

Note that these quantities defined in [Definition B.19](#) are non-negative and finite by [Assumption B.17](#).

Lemma B.3.4. *For any $x_0 \in \mathbb{R}^d$, for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$,*

$$e^{\frac{E(\mathcal{Q}|x_0)-\varepsilon}{\eta}} \leq \mathbb{E}_x[\tilde{\tau}_Q^K] \leq e^{\frac{E(\mathcal{Q}|x_0)+\varepsilon}{\eta}}. \quad (\text{B.3.20})$$

Proof. If $x_0 \in \bigcup_{i=1}^K \mathcal{C}_i^\delta$, then Lemma B.3.3 applies and yields the result. If $x_0 \in \bigcup_{i=K+1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$, then the result holds trivially. Let us now consider the general case where $x_0 \in \mathbb{R}^d \setminus \bigcup_{i=1}^{N_{\text{crit}}} \mathcal{C}_i^\delta$ and let us first prove the upper bound, the lower bound follows similarly.

By Lemma B.3.3 and Lemma B.2.13: for small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there exists $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$, $i \in \{1, \dots, K\}$, $x' \in \mathcal{C}_i^\delta$, both

$$\mathbb{E}_{x'}[\tilde{\tau}_Q^K] \leq \exp\left(\frac{E(\mathcal{Q}|i) + \varepsilon}{\eta}\right) \quad (\text{B.3.21})$$

and,

$$\mathcal{Q}^K_{\mathcal{C}^\delta}(x_0, \mathcal{C}_i^\delta) \leq \exp\left(-\frac{B_{x_0,i}}{\eta} + \frac{\varepsilon}{\eta}\right) \quad (\text{B.3.22})$$

hold.

Using the strong Markov property at time τ_K^η (which is finite almost surely by Corollary B.2.17), we have:

$$\begin{aligned} \mathbb{E}_{x_0}[\tilde{\tau}_Q^K] &= \sum_{i=1}^{N_{\text{crit}}} \mathbb{E}_{x_0} \left[(1 + \mathbb{E}_{x_{\tau_K^\eta}^\eta}[\tilde{\tau}_Q^K]) \mathbb{1} \left\{ x_{\tau_K^\eta}^\eta \in \mathcal{C}_i^\delta \right\} \right] \\ &= 1 + \sum_{i=1}^K \mathbb{E}_{x_0} \left[\mathbb{E}_{x_{\tau_K^\eta}^\eta}[\tilde{\tau}_Q^K] \mathbb{1} \left\{ x_{\tau_K^\eta}^\eta \in \mathcal{C}_i^\delta \right\} \right] + \sum_{i=K+1}^{N_{\text{crit}}} 0 \times \mathbb{1} \left\{ x_{\tau_K^\eta}^\eta \in \mathcal{C}_i^\delta \right\} \\ &\leq 1 + \sum_{i=1}^K \exp\left(\frac{E(\mathcal{Q}|i) - B_{x_0,i} + 2\varepsilon}{\eta}\right), \end{aligned} \quad (\text{B.3.23})$$

where we used (B.3.21) and (B.3.22) in the last inequality. Bounding the sum as follows

$$\begin{aligned} 1 + \sum_{i=1}^K \exp\left(\frac{E(\mathcal{Q}|i) - B_{x_0,i} + 2\varepsilon}{\eta}\right) &\leq (K+1) \exp\left(\frac{0 \vee \max_{i \in \{1, \dots, K\}} (E(\mathcal{Q}|i) - B_{x_0,i} + 2\varepsilon)}{\eta}\right) \\ &= (K+1) \exp\left(\frac{E(\mathcal{Q}|x_0) + 2\varepsilon}{\eta}\right), \end{aligned} \quad (\text{B.3.24})$$

yields the upper bound.

The lower bound follows similarly using Lemma B.2.14 and the lower bound from Lemma B.3.3. ■

Theorem B.5. *For any $x_0 \in \mathbb{R}^d$, for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$,*

$$e^{\frac{E(\mathcal{Q}|x_0)-\varepsilon}{\eta}} \leq \mathbb{E}_{x_0}[\tau_Q^\eta] \leq e^{\frac{E(\mathcal{Q}|x_0)+\varepsilon}{\eta}}. \quad (\text{B.3.25})$$

Proof. The result follows directly by combining Lemma B.3.2 and Lemma B.3.4.

Specifically, by [Lemma B.3.4](#), for any $\varepsilon > 0$, for small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there exists $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$:

$$e^{\frac{E(\mathcal{Q}|x_0)+\varepsilon}{\eta}} \leq \mathbb{E}_{x_0}[\tilde{\tau}_Q^K] \leq e^{\frac{E(\mathcal{Q}|x_0)+\varepsilon}{\eta}}. \quad (\text{B.3.26})$$

Then by [Lemma B.3.2](#), potentially reducing η_0 , we have:

$$\mathbb{E}_{x_0}[\tilde{\tau}_Q^K] \leq \mathbb{E}_{x_0}[\tau_Q^\eta] \leq \mathbb{E}_{x_0}[\tilde{\tau}_Q^K] \times e^{\frac{\varepsilon}{\eta}}. \quad (\text{B.3.27})$$

Combining these inequalities and using the fact that $\varepsilon > 0$ was arbitrary concludes the proof. $\blacksquare$

The following lemma, though looking at first weaker than [Theorem B.5](#), will be useful later due to its uniformity in the initial condition in neighborhoods of the components.

Lemma B.3.6. *Under [Assumption B.17](#), for any $i \in \mathcal{V} \setminus \mathcal{Q}$, for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$, for any $x \in \mathcal{C}_i^\delta$,*

$$e^{\frac{E(\mathcal{Q}|i)-\varepsilon}{\eta}} \leq \mathbb{E}_x[\tau_Q^\eta] \leq e^{\frac{E(\mathcal{Q}|i)+\varepsilon}{\eta}}. \quad (\text{B.3.28})$$

Proof. The result follows directly by combining [Lemma B.3.2](#) and [Lemma B.3.3](#).

Specifically, by [Lemma B.3.3](#), for any $\varepsilon > 0$, for small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there exists $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$, for any $x \in \mathcal{C}_i^\delta$:

$$e^{\frac{E(\mathcal{Q}|i)+\varepsilon}{\eta}} \leq \mathbb{E}_x[\tilde{\tau}_Q^K] \leq e^{\frac{E(\mathcal{Q}|i)+\varepsilon}{\eta}}. \quad (\text{B.3.29})$$

Then by [Lemma B.3.2](#), potentially reducing η_0 , we have:

$$\mathbb{E}_x[\tilde{\tau}_Q^K] \leq \mathbb{E}_x[\tau_Q^\eta] \leq \mathbb{E}_x[\tilde{\tau}_Q^K] \times e^{\frac{\varepsilon}{\eta}}. \quad (\text{B.3.30})$$

Combining these inequalities and using the fact that $\varepsilon > 0$ was arbitrary concludes the proof. $\blacksquare$

Another technical lemma that will be useful later is the following.

Lemma B.3.7. *For any $x_0 \in \mathbb{R}^d$, for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$,*

$$e^{\frac{E(\mathcal{Q}|x_0)-\varepsilon}{\eta}} \leq \mathbb{E}_{x_0}[\tilde{\tau}_Q] \quad (\text{B.3.31})$$

Proof. This result follows directly from [Lemma B.3.4](#) and the fact that $\tilde{\tau}_Q \leq \tilde{\tau}_Q^K$. $\blacksquare$

B.3.3 Hitting time of SGD

Leveraging the bounds on the hitting time of the accelerated process, we can now derive bounds on the hitting time of SGD. Denote by τ the hitting time of $\mathcal{C}^\delta = \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$ for the SGD sequence:

$$\tau := \inf \left\{ n \geq 0 : x_n \in \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta \right\}. \quad (\text{B.3.32})$$

It corresponds to the hitting time of SGD defined in the main text (B.3.32) in the simpler case where each $\mathcal{C}_j^\delta$ is of the form $\mathcal{C}_j^\delta = \mathcal{U}_\delta(\mathcal{C}_j)$ for $j \in \mathcal{Q}$. We now state our first main results on the actual SGD sequence (B.1.12). The theorem below is a corollary of Theorem B.5 which concerned the subsampled process (B.1.16).

Theorem B.8. *For any $x_0 \in \mathbb{R}^d$, for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$,*

$$\mathbb{E}_{x_0}[\tau] \leq e^{\frac{E(\mathcal{Q}|x_0)+\varepsilon}{\eta}}. \quad (\text{B.3.33})$$

Proof. This result follows from Theorem B.5 and the fact that, by construction,

$$\tau \leq \tau_{\mathcal{Q}}^\eta \lceil \eta^{-1} \rceil, \quad (\text{B.3.34})$$

with $\lceil \eta^{-1} \rceil \leq e^{\frac{\varepsilon}{\eta}}$ for small enough η . $\blacksquare$

Corollary B.3.9. *For any $x_0 \in \mathbb{R}^d$, for any $\Delta > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$,*

$$\mathbb{P}_x \left(\tau < \exp \left(\frac{E(\mathcal{Q}|x_0) + \Delta}{\eta} \right) \right) \leq e^{-\frac{\Delta}{2\eta}}. \quad (\text{B.3.35})$$

Proof. By Markov's inequality and Theorem B.8, for any $x_0 \in \mathbb{R}^d$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there exists $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$:

$$\mathbb{P}_x \left(\tau \geq \exp \left(\frac{E(\mathcal{Q}|x_0) + \Delta}{\eta} \right) \right) \leq \frac{\mathbb{E}_x[\tau]}{\exp \left(\frac{E(\mathcal{Q}|x_0) + \Delta}{\eta} \right)} \quad (\text{B.3.36})$$

$$\leq \frac{\exp \left(\frac{E(\mathcal{Q}|x_0) + \Delta/2}{\eta} \right)}{\exp \left(\frac{E(\mathcal{Q}|x_0) + \Delta}{\eta} \right)} = \exp(-\Delta/2\eta). \quad (\text{B.3.37})$$

$\blacksquare$

Let us now focus on the lower-bound which requires more care. We will require the following additional assumption:

Assumption B.20. Assume that there exist $\mu > 0$, $\bar{\sigma}_\infty^2 < +\infty$, $R > 0$, such that,

1. $\text{cl} \mathcal{U}_R(\mathcal{Q}_j)$, $j = 1, \dots, N_{\text{targ}}$ are pairwise disjoint.
2. for all $j = 1, \dots, N_{\text{targ}}$, $x \in \text{cl} \mathcal{U}_R(\mathcal{Q}_j)$, $U(x, \omega)$ is a $\bar{\sigma}_\infty^2$ -sub-Gaussian:

$$\forall p \in \mathbb{R}^d, \quad \log \mathbb{E} \left[e^{\langle p, U(x, \omega) \rangle} \right] \leq \frac{\bar{\sigma}_\infty^2}{2} \|p\|^2. \quad (\text{B.3.38})$$

3. for all $j = 1, \dots, N_{\text{targ}}$, $x \in \text{cl} \mathcal{U}_R(\mathcal{Q}_j)$, there exists $x' \in \text{proj}_{\mathcal{Q}_j}(x)$ a projection of x on $\mathcal{Q}_j$ such that:

$$\langle \nabla f(x), x - x' \rangle \geq \frac{\mu}{2} \|x - x'\|^2 \quad (\text{B.3.39})$$

Note that, by [Assumption B.3](#) and compactness of $\text{crit}(f)$, such a $\bar{\sigma}_\infty^2$ always exists since one can take

$$\bar{\sigma}_\infty^2 = \sup \left\{ \sigma_\infty^2(f(x)) : x \in \bigcup_{j=1}^{N_{\text{targ}}} \text{cl} \mathcal{U}_R(\mathcal{Q}_j) \right\}. \quad (\text{B.3.40})$$

In particular, if σ_∞^2 is constant, then one can simply take $\bar{\sigma}_\infty^2 = \sigma_\infty^2$.

Lemma B.3.10. *Under [Assumption B.20](#), for any $j \in \mathcal{Q}$, there exists $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$, for any $x_0 \in \mathbb{R}^d$ satisfying $d(x_0, \mathcal{Q}_j) \leq \frac{R}{12}$, we have, for SGD started at x_0 :*

$$\mathbb{P}_{x_0}(\forall n \geq 0, d(x_n, \mathcal{Q}_j) \leq R) \geq 1 - \exp\left(-\frac{\mu R^2}{1152 \bar{\sigma}_\infty^2 \eta}\right). \quad (\text{B.3.41})$$

Proof. Let us denote by $\mathcal{K} := \text{cl} \mathcal{U}_R(\mathcal{Q}_j)$. Given $x_0 \in \mathcal{K}$, we define the projected SGD sequence as

$$\begin{cases} p_0 = x_0 \\ p_{n+1} \in \text{proj}_{\mathcal{K}}(p_n - \eta \nabla f(p_n) + \eta U_n), \end{cases} \quad \text{where } U_n = U(p_n, \omega_n). \quad (\text{B.3.42})$$

By [Assumption B.20](#), for any $n \geq 0$, there exists $y_n \in \text{proj}_{\mathcal{Q}_j}(p_n)$ such that:

$$\langle \nabla f(p_n), p_n - y_n \rangle \geq \frac{\mu}{2} \|p_n - y_n\|^2. \quad (\text{B.3.43})$$

Define the martingale sequence $(M_n)_{n \geq 0}$ by:

$$M_n := \sum_{k=0}^n (1 - \mu\eta)^{n-k} \langle U_k, p_k - y_k \rangle. \quad (\text{B.3.44})$$

We recursively compute the moment generating function of M_n : for $\alpha \geq 0$, $n \geq 0$, by the tower property of conditional expectation and [Assumption B.20](#):

$$\begin{aligned} \mathbb{E}[e^{\alpha M_n}] &\leq \exp\left(\frac{\alpha^2 R^2 \bar{\sigma}_\infty^2}{2}\right) \exp\left(\frac{(1 - \mu\eta)^2 \alpha^2 \bar{\sigma}_\infty^2}{2}\right) \dots \exp\left(\frac{(1 - \mu\eta)^{2n} \alpha^2 \bar{\sigma}_\infty^2}{2}\right) \\ &= \exp\left(\frac{\alpha^2 R^2 \bar{\sigma}_\infty^2}{2} \times \frac{1 - (1 - \mu\eta)^{2(n+1)}}{1 - (1 - \mu\eta)^2}\right) \\ &\leq \exp\left(\frac{\alpha^2 R^2 \bar{\sigma}_\infty^2}{2\mu\eta}\right), \end{aligned} \quad (\text{B.3.45})$$

provided that η is small enough so that $\mu\eta < 1$. Note that we used that the iterates of [\(B.3.42\)](#) are at distance at most R from $\mathcal{Q}_j$ by construction.

Since $(e^{\alpha M_n})_{n \geq 0}$ is a convex function of the martingale sequence $(M_n)_{n \geq 0}$, it is a sub-martingale. We can apply Doob's maximal inequality to the non-negative sub-martingale $(M_n)_{n \geq 0}$ to obtain that for any $\Delta > 0$, $N \geq 0$:

$$\begin{aligned} \mathbb{P}\left(\sup_{n \leq N-1} M_n \geq \Delta\right) &\leq \mathbb{P}\left(\sup_{n \leq N-1} e^{\alpha M_n} \geq e^{\alpha \Delta}\right) \\ &\leq \exp\left(\frac{\alpha^2 R^2 \bar{\sigma}_\infty^2}{2\mu\eta} - \alpha \Delta\right), \end{aligned} \quad (\text{B.3.46})$$

where we used (B.3.45) in the last inequality. Optimizing the right-hand side of (B.3.46) with respect to α and setting $\alpha = \frac{\mu\eta\Delta}{R^2\bar{\sigma}_\infty^2}$, we obtain:

$$\mathbb{P}\left(\sup_{n \leq N-1} M_n \geq \Delta\right) \leq \exp\left(-\frac{\mu\eta\Delta^2}{2R^2\bar{\sigma}_\infty^2}\right). \quad (\text{B.3.47})$$

Taking $\Delta \leftarrow \Delta/\eta$ and $N \rightarrow +\infty$, we obtain that, by monotone continuity of probability measures, for any $\Delta > 0$:

$$\mathbb{P}\left(\sup_{n \geq 0} \eta M_n \geq \Delta\right) \leq \exp\left(-\frac{\mu\Delta^2}{2\eta R^2\bar{\sigma}_\infty^2}\right). \quad (\text{B.3.48})$$

Denote by $G := \sup_{x \in \mathcal{K}, \omega \in \Omega} \|\nabla f(x) + U(x, \omega)\|$ the bound on the gradient and on the noise on $\mathcal{K}$, which is finite by [Assumption B.1](#) and [Assumption B.3](#).

Let us now derive a recursive inequality for the distance to $\mathcal{Q}_j$. For any iterate $n \geq 0$, by non-expansiveness of the projection, we have:

$$\begin{aligned} d^2(p_{n+1}, \mathcal{Q}_j) &\leq \|p_{n+1} - y_n\|^2 \\ &\leq \|p_n - \eta \nabla f(p_n) + \eta U_n - y_n\|^2 \\ &= \|p_n - y_n\|^2 + \eta^2 \|\nabla f(p_n) - U_n\|^2 \\ &\quad - 2\eta \langle \nabla f(p_n), p_n - y_n \rangle + 2\eta \langle U_n, p_n - y_n \rangle \\ &\leq (1 - \mu\eta) \|p_n - y_n\|^2 + \eta^2 G^2 + 2\eta \langle U_n, p_n - y_n \rangle, \end{aligned} \quad (\text{B.3.49})$$

where we used (B.3.43) in the last inequality.

Iterating this inequality and using that $d(p_n, \mathcal{Q}_j) = \|p_n - y_n\|$, we obtain:

$$\begin{aligned} d^2(p_n, \mathcal{Q}_j) &\leq (1 - \mu\eta)^n d^2(x_0, \mathcal{Q}_j) + \eta^2 G^2 \sum_{i=0}^{n-1} (1 - \mu\eta)^i \\ &\quad + 2\eta \sum_{i=0}^{n-1} (1 - \mu\eta)^{n-1-i} \langle U_i, p_i - y_i \rangle \\ &= (1 - \mu\eta)^n d^2(x_0, \mathcal{Q}_j) + \eta^2 G^2 \frac{1 - (1 - \mu\eta)^n}{\mu\eta} + 2\eta M_{n-1} \\ &\leq (1 - \mu\eta)^n d^2(x_0, \mathcal{Q}_j) + \frac{2\eta G^2}{\mu} + 2\eta M_{n-1}. \end{aligned} \quad (\text{B.3.50})$$

Using that $d(x_0, \mathcal{Q}_j) \leq \frac{R}{12}$ by assumption, requiring that $\eta > 0$ be small enough so that $\frac{2\eta G^2}{\mu} \leq \frac{R^2}{12}$ and taking $\Delta \leftarrow \frac{R^2}{24}$ in (B.3.48), we obtain that for any $n \geq 0$, with probability at least $1 - \exp\left(-\frac{\mu R^2}{1152\bar{\sigma}_\infty^2\eta}\right)$:

$$d^2(p_n, \mathcal{Q}_j) \leq \frac{R^2}{4}. \quad (\text{B.3.51})$$

In addition, take $\eta > 0$ small enough so that $\frac{R}{2} + \eta G \leq \frac{R}{2}$: this implies that if $x \in \mathbb{R}^d$ so that $d(x, \mathcal{Q}_j) \leq \frac{R}{2}$, then $x - \eta \nabla f(x) + \eta U(x, \omega) \in \mathcal{K}$. Combining this remark with (B.3.51), we can show recursively that the sequences $(x_n)_{n \geq 0}$ and $(p_n)_{n \geq 0}$ coincide with probability at least $1 - \exp\left(-\frac{\mu R^2}{1152\bar{\sigma}_\infty^2\eta}\right)$, yielding the desired result. ■

Define by $\tilde{\tau}_{\mathcal{Q}}$ the hitting time of $\bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$ by the induced chain $(z_n)_{n \geq 0}$ (see [Definition B.12](#)).

Lemma B.3.11. *Define*

$$\mathcal{B}_\infty := \max_{i \in \mathcal{V} \setminus \mathcal{Q}} \min_{j \in \mathcal{Q}} \mathcal{B}_{i,j} + 1. \quad (\text{B.3.52})$$

For any $x_0 \in \mathbb{R}^d$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$, for any $n \geq 0$,

$$\mathbb{P}_{x_0}(\tilde{\tau}_{\mathcal{Q}} > n) \leq \left(1 - e^{-\frac{\mathcal{B}_\infty}{\eta}}\right)^n. \quad (\text{B.3.53})$$

Note that [Assumption B.17](#) ensures that $\mathcal{B}_\infty$ is finite.

Proof. Fix some $x_0 \in \mathbb{R}^d$. By definition of $\mathcal{B}_\infty$, for any vertex $i \in \mathcal{V} \setminus \mathcal{Q}$, there exists $j \in \mathcal{Q}$ such that:

$$\mathcal{B}_{i,j} \leq \mathcal{B}_\infty - 1. \quad (\text{B.3.54})$$

Fix $\varepsilon \leftarrow 1$. By [Lemma B.2.4](#), for small enough neighborhoods $\mathcal{C}_i^\delta$ of $\mathcal{C}_i$, $i = 1, \dots, N_{\text{crit}}$, there exists $\eta_0 > 0$ such that for all $i \in \mathcal{V} \setminus \mathcal{Q}$, $x \in \mathcal{C}_i^\delta$, $0 < \eta < \eta_0$:

$$\begin{aligned} \mathbb{Q}_{\mathcal{C}^\delta} \left(x, \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta \right) &\geq \max_{j \in \mathcal{Q}} \mathbb{Q}_{\mathcal{C}^\delta}(x, \mathcal{C}_j^\delta) \\ &\geq \exp \left(-\frac{\min_{j \in \mathcal{Q}} \mathcal{B}_{i,j}}{\eta} - \frac{\varepsilon}{\eta} \right) \\ &\geq \exp \left(-\frac{\mathcal{B}_\infty}{\eta} \right). \end{aligned} \quad (\text{B.3.55})$$

Therefore, for any $n \geq 1$, by the strong Markov property, we have:

$$\begin{aligned} \mathbb{P}_{x_0}(\tilde{\tau}_{\mathcal{Q}} > n) &= \mathbb{E}_{x_0} \left[\mathbb{1}\{\tilde{\tau}_{\mathcal{Q}} > n-1\} \mathbb{1}\{z_n \notin \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta\} \right] \\ &= \mathbb{E}_{x_0} \left[\mathbb{1}\{\tilde{\tau}_{\mathcal{Q}} > n-1\} \mathbb{E}_{z_{n-1}} \left[\mathbb{1}\{z_n \notin \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta\} \right] \right] \\ &\leq \left(1 - \exp \left(-\frac{\mathcal{B}_\infty}{\eta} \right) \right) \mathbb{P}_{x_0}(\tilde{\tau}_{\mathcal{Q}} > n-1), \end{aligned} \quad (\text{B.3.56})$$

where we used the (B.3.55) in the last inequality. Iterating (B.3.56), we obtain the desired result. $\blacksquare$

Theorem B.12. Fix $x_0 \in \mathbb{R}^d$. Under [Assumption B.20](#) and assuming that,

$$\frac{\mu R^2}{1152 \bar{\sigma}_\infty^2} > \mathcal{B}_\infty - E(\mathcal{Q} | x_0), \quad (\text{B.3.57})$$

for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$:

$$\mathbb{E}_{x_0}[\tau] \geq e^{\frac{E(\mathcal{Q} | x_0) - \varepsilon}{\eta}}. \quad (\text{B.3.58})$$

Remark 7. Note that the condition (B.3.57) is implied by the stronger condition

$$\frac{\mu R^2}{1152\bar{\sigma}_\infty^2} > \mathcal{B}_\infty. \quad (\text{B.3.59})$$

In particular, $\mathcal{B}_\infty$ does not depend on f nor on the noise distribution on the interior of the basins of attraction of $\mathcal{Q}_1, \dots, \mathcal{Q}_{N_{\text{targ}}}$. Indeed, by definition, $\mathcal{B}_\infty$ is equal to

$$\mathcal{B}_\infty = \max_{i \in \mathcal{V} \setminus \mathcal{Q}} \min_{j \in \mathcal{Q}} \mathcal{B}_{i,j} + 1, \quad (\text{B.3.60})$$

and, for $i \in \mathcal{V} \setminus \mathcal{Q}, j \in \mathcal{Q}$, thanks to Lemma B.1.7, we have that

$$\mathcal{B}_{i,j} = \inf \left\{ \mathcal{S}_T(\gamma) : \gamma \in \mathcal{C}([0, T]) \gamma_0 \in \mathcal{C}_i, \gamma_T \in \text{Attr}(\mathcal{C}_j), \gamma_n \notin \bigcup_{k \in \mathcal{V}} \mathcal{C}_k \forall n \in 1, \dots, T-1 \right\}, \quad (\text{B.3.61})$$

since for any $x \in \text{Attr}(\mathcal{C}_j)$, the gradient flow started at x converges to $\mathcal{C}_j$ and has zero action cost (Lemma B.1.5).

Hence, this condition assumption will be satisfied for functions with a sufficiently sharp profile near $\mathcal{Q}_1, \dots, \mathcal{Q}_{N_{\text{targ}}}$ and a given profile away from it.

Proof. Let us first begin by invoking Lemma B.3.7: for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$,

$$e^{\frac{\mathbb{E}(\mathcal{Q} | x_0) - \varepsilon}{\eta}} \leq \mathbb{E}_{x_0}[\tilde{\tau}_{\mathcal{Q}}]. \quad (\text{B.3.62})$$

Also require that, for all $j \in \mathcal{Q}$, $\mathcal{C}_j^\delta \subset \{x \in \mathbb{R}^d : d(x, \mathcal{Q}_j) \leq \frac{R}{12}\}$ with R as in Assumption B.20.

If $x_0 \in \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$, then the result is trivial so let us assume that $x_0 \notin \bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$.

Let us denote by

$$H := \left\{ \forall n \geq 0, d(x_{n+\tau}, \mathcal{Q}_j) \leq R \text{ with } x_\tau \in \mathcal{C}_j^\delta \right\} \quad (\text{B.3.63})$$

the event that, after SGD hits $\bigcup_{k \in \mathcal{Q}} \mathcal{C}_k^\delta$ at $\mathcal{C}_j^\delta$, it stays at distance at most R from $\mathcal{Q}_j$ for all subsequent iterates. Note that Theorem B.8 ensures that τ is finite almost surely. By the strong Markov property, we thus obtain from Lemma B.3.10 that

$$\mathbb{P}_{x_0}(H) \geq 1 - \exp\left(-\frac{\mu R^2}{1152\bar{\sigma}_\infty^2 \eta}\right). \quad (\text{B.3.64})$$

Let us now start from $\mathbb{E}_{x_0}[\tilde{\tau}_{\mathcal{Q}}]$ where $\tilde{\tau}_{\mathcal{Q}}$ is the hitting time of $\bigcup_{j \in \mathcal{Q}} \mathcal{C}_j^\delta$ by the induced chain $(z_n)_{n \geq 0}$ (see Definition B.12), for which we have a lower-bound from Lemma B.3.7. We write

$$\mathbb{E}_{x_0}[\tilde{\tau}_{\mathcal{Q}}] = \mathbb{E}_{x_0}[\tilde{\tau}_{\mathcal{Q}} \mathbb{1}_H] + \mathbb{E}_{x_0}[\tilde{\tau}_{\mathcal{Q}} \mathbb{1}_{H^c}]. \quad (\text{B.3.65})$$

We begin with the first term of the RHS of (B.3.65). With the notations from Definition B.12, let us consider $k \geq 0$ such that $\tau_k^\eta \lfloor \eta^{-1} \rfloor < \tau \leq \tau_{k+1}^\eta \lfloor \eta^{-1} \rfloor$ (which is possible since these hitting times are finite almost surely and $\tau > 0$). But,

on the event H , $z_{k+1} = x_{\tau_{k+1}}^\eta = x_{\tau_{k+1}^\eta \lfloor \eta^{-1} \rfloor}^\eta$ must be in $\mathcal{C}_j^\delta$. This means that, on H , $\tilde{\tau}_Q = k + 1$. Moreover, $\tau \leq k \lfloor \eta^{-1} \rfloor$ so that, on H

$$\tilde{\tau}_Q \leq \frac{\tau}{\lfloor \eta^{-1} \rfloor} + 1. \quad (\text{B.3.66})$$

and,

$$\mathbb{E}_{x_0}[\tilde{\tau}_Q \mathbb{1}_H] \leq \mathbb{E}_{x_0} \left[\frac{\tau}{\lfloor \eta^{-1} \rfloor} + 1 \right]. \quad (\text{B.3.67})$$

Let us now focus on the second term of the RHS of (B.3.65). By Fubini's theorem for non-negative integrands, we can rewrite the expectation as:

$$\begin{aligned} \mathbb{E}_{x_0}[\tilde{\tau}_Q \mathbb{1}_{H^c}] &= \mathbb{E}_{x_0} \left[\int_0^{+\infty} \mathbb{1}_{\{\tilde{\tau}_Q > t\}} \mathbb{1}_{H^c} dt \right] \\ &= \int_0^{+\infty} \mathbb{P}_{x_0}(\tilde{\tau}_Q > t, H^c) dt. \end{aligned} \quad (\text{B.3.68})$$

Let us now split the integral into two parts. For $T > 0$, we have:

$$\begin{aligned} \mathbb{E}_{x_0}[\tilde{\tau}_Q \mathbb{1}_{H^c}] &= \int_0^T \mathbb{P}_{x_0}(\tilde{\tau}_Q > t, H^c) dt + \int_T^{+\infty} \mathbb{P}_{x_0}(\tilde{\tau}_Q > t, H^c) dt \\ &\leq \int_0^T \mathbb{P}_{x_0}(H^c) dt + \int_T^{+\infty} \mathbb{P}_{x_0}(\tilde{\tau}_Q > t) dt. \end{aligned} \quad (\text{B.3.69})$$

By (B.3.64), the first term of the RHS of (B.3.69) is upper-bounded as:

$$\int_0^T \mathbb{P}_{x_0}(H^c) dt \leq T \exp \left(-\frac{\mu R^2}{1152 \bar{\sigma}_\infty^2 \eta} \right). \quad (\text{B.3.70})$$

By (B.3.57), take $\varepsilon > 0$ small enough so that

$$\frac{\mu R^2}{1152 \bar{\sigma}_\infty^2} \geq \mathcal{B}_\infty - E(Q | x_0) + 3\varepsilon, \quad (\text{B.3.71})$$

and define

$$T := \exp \left(\frac{\mu R^2}{1152 \bar{\sigma}_\infty^2 \eta} + \frac{E(Q | x_0) - 2\varepsilon}{\eta} \right). \quad (\text{B.3.72})$$

(B.3.70) now becomes:

$$\int_0^T \mathbb{P}_{x_0}(H^c) dt \leq e^{\frac{E(Q | x_0) - 2\varepsilon}{\eta}} \quad (\text{B.3.73})$$

By Lemma B.3.11, the second term of the RHS of (B.3.69) can be upper-bounded as:

$$\begin{aligned} \int_T^{+\infty} \mathbb{P}_{x_0}(\tilde{\tau}_Q > t) dt &\leq \int_T^{+\infty} \left(1 - e^{-\frac{\mathcal{B}_\infty}{\eta} t} \right) dt \\ &\leq \int_T^{+\infty} \exp \left(e^{-\frac{\mathcal{B}_\infty}{\eta} t} \right) dt \\ &= e^{\frac{\mathcal{B}_\infty}{\eta}} \exp \left(-\frac{\mathcal{B}_\infty}{\eta} T \right), \end{aligned} \quad (\text{B.3.74})$$

where we used $1 - x \leq e^{-x}$ for any $x \in \mathbb{R}$ in the last inequality. Injecting the definition of T (B.3.72), this bound becomes

$$\begin{aligned} \int_T^{+\infty} \mathbb{P}_{x_0}(\tilde{\tau}_{\mathcal{Q}} > t) dt &\leq \exp\left(\frac{\mathcal{B}_{\infty}}{\eta} - \exp\left(\frac{\mu R^2}{1152\sigma_{\infty}^2\eta} + \frac{E(\mathcal{Q} | x_0) - \mathcal{B}_{\infty} - 2\varepsilon}{\eta}\right)\right) \\ &\leq \exp\left(\frac{\mathcal{B}_{\infty}}{\eta} - \exp\left(\frac{\varepsilon}{\eta}\right)\right), \end{aligned} \quad (\text{B.3.75})$$

where we used (B.3.71) in the last inequality. With $\eta > 0$ small enough, we finally obtain

$$\int_T^{+\infty} \mathbb{P}_{x_0}(\tilde{\tau}_{\mathcal{Q}} > t) dt \leq 1. \quad (\text{B.3.76})$$

Combining (B.3.73) and (B.3.76) in (B.3.69), we obtain:

$$\mathbb{E}_{x_0}[\tilde{\tau}_{\mathcal{Q}} \mathbf{1}_{H^c}] \leq e^{\frac{E(\mathcal{Q} | x_0) - 2\varepsilon}{\eta}} + 1, \quad (\text{B.3.77})$$

and, combining (B.3.67) and (B.3.77) above in (B.3.65), we get:

$$\mathbb{E}_{x_0}[\tilde{\tau}_{\mathcal{Q}}] \leq \mathbb{E}_{x_0}\left[\frac{\tau}{\lfloor \eta^{-1} \rfloor} + 1\right] + e^{\frac{E(\mathcal{Q} | x_0) - 2\varepsilon}{\eta}} + 1. \quad (\text{B.3.78})$$

Plugging in (B.3.62) yields:

$$e^{\frac{E(\mathcal{Q} | x_0) - \varepsilon}{\eta}} \leq \mathbb{E}_{x_0}\left[\frac{\tau}{\lfloor \eta^{-1} \rfloor} + 1\right] + e^{\frac{E(\mathcal{Q} | x_0) - 2\varepsilon}{\eta}} + 1, \quad (\text{B.3.79})$$

which yields the desired result by taking $\eta > 0$ small enough. ■

Let us state a variant of [Theorem B.12](#) that will be useful later.

Lemma B.3.13. Fix $i \in \{1, \dots, N_{\text{crit}}\}$. Under [Assumption B.20](#) and assuming that,

$$\frac{\mu R^2}{1152\sigma_{\infty}^2} > \mathcal{B}_{\infty} - E(\mathcal{Q} | i), \quad (\text{B.3.80})$$

for any $\varepsilon > 0$, for $\mathcal{C}_1^{\delta}, \dots, \mathcal{C}_{N_{\text{crit}}}^{\delta}$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$ such that for all $0 < \eta < \eta_0$, $x \in \mathcal{C}_i^{\delta}$:

$$\mathbb{E}_x[\tau] \geq e^{\frac{E(\mathcal{Q} | i) - \varepsilon}{\eta}}. \quad (\text{B.3.81})$$

The only difference with [Theorem B.12](#) is that the result is uniform over the initial state in $\mathcal{C}_i^{\delta}$ instead of being specific to a given initial point. Its proof is identical to the proof of [Theorem B.12](#): it simply uses [Lemma B.3.6](#) to obtain the necessary variant of [Lemma B.3.7](#) and then follows the same steps. It is therefore omitted.

B.3.4 Basin-dependent convergence result

Let us recall the potential function from ([Azizian et al., 2024b](#)).

Definition B.21 (Potential, ([Azizian et al., 2024b](#), Def. 4)). Define, for $x \in \mathbb{R}^d$

$$V_{\infty}(x) = 2\alpha_{\infty} \circ f(x) \quad (\text{B.3.82})$$

where $\alpha_\infty : \mathbb{R}^d \rightarrow \mathbb{R}$ is a twice continuously differentiable primitive of $1/\sigma_\infty^2$.

Lemma B.3.14. *For any $\gamma \in \mathcal{C}([0, T])$, $t \in [0, T]$, we have*

$$V_\infty(\gamma(t)) - V_\infty(\gamma(0)) \leq \mathcal{S}_{0,T}(\gamma). \quad (\text{B.3.83})$$

Proof. We have that, with V_∞ defined in [Definition B.21](#), by Young's inequality,

$$\begin{aligned} V_\infty(\gamma_t) - V_\infty(\gamma_0) &= 2 \int_0^t \frac{\langle \dot{\gamma}_s, \nabla f(\gamma_s) \rangle}{\sigma_\infty^2 \circ f(\gamma_s)} ds \\ &\leq \int_0^t \frac{\|\dot{\gamma}_s\|^2}{2\sigma_\infty^2 \circ f(\gamma_s)} ds + \int_0^t \frac{\|\nabla f(\gamma_s)\|^2}{2\sigma_\infty^2 \circ f(\gamma_s)} ds + \int_0^t \frac{\langle \dot{\gamma}_s, \nabla f(\gamma_s) \rangle}{\sigma_\infty^2 \circ f(\gamma_s)} ds \\ &= \int_0^t \frac{\|\dot{\gamma}_s + \nabla f(\gamma_s)\|^2}{2\sigma_\infty^2 \circ f(\gamma_s)} ds \\ &\leq \int_0^t \mathcal{L}(\gamma_s, \dot{\gamma}_s) ds \\ &\leq \mathcal{S}_{0,T}(\gamma). \end{aligned} \quad (\text{B.3.84})$$

where we used [Lemma B.1.2](#) in the last inequality. $\blacksquare$

Lemma B.3.15. *V_∞ is coercive i.e., $V_\infty(x) \rightarrow +\infty$ as $\|x\| \rightarrow +\infty$.*

Proof. By [Assumption B.4](#), $\sigma_\infty^2(s) = o(s^2)$ as $s \rightarrow \infty$ and therefore α a primitive of $1/\sigma_\infty^2$ must be coercive. By coercivity of f ([Assumption B.1](#)), $V_\infty = 2\alpha \circ f$ is also coercive. $\blacksquare$

Lemma B.3.16. *For any $\mathcal{K} \subset \mathbb{R}^d$ compact, for any $c > 0$, the set*

$$\{\gamma(t) : t \in [0, T], \gamma \in \mathcal{C}([0, T]), T \in \mathbb{N}, \gamma_0 \in \mathcal{K}, \mathcal{S}_{t,T}(\gamma) \leq c\} \quad (\text{B.3.85})$$

is included in the set

$$\left\{x \in \mathbb{R}^d : V_\infty(x) \leq c + \sup_{\mathcal{K}} V_\infty\right\}, \quad (\text{B.3.86})$$

which is compact.

Proof. Consider such a path $\gamma \in \mathcal{C}([0, T])$ with $T \in \mathbb{N}$. By [Lemma B.3.14](#), we have, for any $t \in [0, T]$,

$$V_\infty(\gamma(t)) - V_\infty(\gamma(0)) \leq \mathcal{S}_{0,T}(\gamma) \leq c. \quad (\text{B.3.87})$$

Hence the set from [\(B.3.85\)](#) is included in

$$\left\{x \in \mathbb{R}^d : V_\infty(x) \leq c + \sup_{\mathcal{K}} V_\infty\right\}. \quad (\text{B.3.88})$$

Now, by [Lemma B.3.15](#), the set of [\(B.3.88\)](#) is compact, so that the set of [\(B.3.85\)](#) is bounded. $\blacksquare$

The following lemma is a slight refinement of ([Azizian et al., 2024b](#), Lem. D.33). For any $C \subset \text{crit}(f)$ connected component of the set of critical points, we define its basin of attraction as the set of points from which the gradient flow converges to C :

$$\text{Attr}(C) := \left\{x \in \mathbb{R}^d : \lim_{t \rightarrow +\infty} d(\Theta_t(x), C) = 0\right\}. \quad (\text{B.3.89})$$

Lemma B.3.17. *For any $C \subset \text{crit}(f)$ connected component which is minimizing, for $\delta > 0$ small enough, for $\delta' \in [0, \delta)$ small enough,*

$$\mathcal{B}(\mathcal{U}'_{\delta}(C), \mathbb{R}^d \setminus \mathcal{U}_{\delta}(C)) > 0. \quad (\text{B.3.90})$$

Moreover, $\mathcal{U}_{\delta}(C)$ can be assumed to be contained in $\text{Attr}(C)$.

Proof. Since C is minimizing, there exists $\delta_0 > 0$ such that, for any $x \in \mathcal{U}_{\delta_0}(C) \setminus C$, $f(x) > f_C$ where f_C is the value of f on C . Moreover, by [Lemma B.1.10](#), taking $\delta > 0$ small enough also ensures that $\mathcal{U}_{\delta_0}(C) \subset \text{Attr}(C)$.

Take $\delta \leq \delta_0$, $\mathcal{U} := \mathcal{U}_{\delta}(C)$ and $\Delta := \min\{V_{\infty}(x) - \alpha_{\infty}(f_C) : x \in \mathbb{R}^d, d(x, C) = \delta/2\}$. By the continuity of V_{∞} and the fact that α_{∞} is (strictly) increasing, we have that $\Delta > 0$.

Finally, take $\delta' \in [0, \delta/2)$ small enough so that, for any $x \in \mathcal{U}_{\delta'}(C)$, $f(x) \leq \alpha_{\infty}(f_C) + \Delta/2$. To conclude the proof of this lemma, we now show that $\mathcal{B}(\mathcal{U}'_{\delta}(C), \mathbb{R}^d \setminus \mathcal{U}_{\delta}(C)) > 0$. Consider some $T > 0$ and $\gamma \in \mathcal{C}([0, T], \mathbb{R}^d)$ such that $\gamma_0 \in \mathcal{U}'_{\delta}(C)$, $\gamma_T \in \mathbb{R}^d \setminus \mathcal{U}_{\delta}(C)$. By continuity of γ and $d(\cdot, C)$, there exists $t \in [0, T]$ such that $d(\gamma_t, C) = \delta/2$. By [Lemma B.3.14](#), we have that

$$\begin{aligned} \Delta &\leq V_{\infty}(\gamma_t) - \alpha_{\infty}(f_C) \\ &= V_{\infty}(\gamma_t) - V_{\infty}(\gamma_0) + V_{\infty}(\gamma_0) - \alpha_{\infty}(f_C) \\ &\leq S_{0,T}(\gamma) + \frac{\Delta}{2}. \end{aligned} \quad (\text{B.3.91})$$

Since this is valid for any γ , we obtain that $\mathcal{B}(\mathcal{U}'_{\delta}(C), \mathbb{R}^d \setminus \mathcal{U}_{\delta}(C)) \geq \Delta/2 > 0$. ■

We now proceed to the key result of this section.

Lemma B.3.18. *For any $C \subset \text{crit}(f)$ connected component which is minimizing, for any $x \in \text{Attr}(C)$,*

$$\mathcal{B}(x, \mathbb{R}^d \setminus \text{Attr}(C)) > 0. \quad (\text{B.3.92})$$

Proof. First, let us invoke [Lemma B.3.17](#) to obtain some $0 < \delta' < \delta$ such that

$$\mathcal{B}(\mathcal{U}'_{\delta'}(C), \mathbb{R}^d \setminus \mathcal{U}_{\delta'}(C)) > 0. \quad (\text{B.3.93})$$

and $\mathcal{U}_{\delta'}(C) \subset \text{Attr}(C)$.

Then, by [Lemma B.3.16](#) and continuity of f and σ_{∞}^2 , the set

$$\{\sigma_{\infty}^2(f(\gamma(t))) : t \in [0, T], \gamma \in \mathcal{C}([0, T]), T \in \mathbb{N}, \gamma_0 = x, S_{t,T}(\gamma) \leq 1\} \quad (\text{B.3.94})$$

is bounded by some finite constant $a \in (0, +\infty)$.

With $(\Theta_t(x))_{t \geq 0}$ the gradient flow starting from x ([Definition B.8](#)), by definition of $\text{Attr}(C)$, there exists $S > 0$ such that

$$\Theta_S(x) \in \mathcal{U}_{\delta/2}(C). \quad (\text{B.3.95})$$

Now, consider some $T \in \mathbb{N}$ and $\gamma \in \mathcal{C}([0, T])$ such that $\gamma_0 = x$, $\gamma_T \in \mathbb{R}^d \setminus \text{Attr}(C)$ and $S_T(\gamma) \leq 1$.

We now define

$$g_t := \frac{1}{2} \|\gamma_t - \Theta_t(x)\|^2. \quad (\text{B.3.96})$$

In particular, γ must be differentiable almost everywhere and so we have

$$\begin{aligned} \dot{g}_t &= \langle \dot{\gamma}_t - \dot{\Theta}_t(x), \gamma_t - \Theta_t(x) \rangle \\ &= \langle \nabla f(\Theta_t(x)) - \nabla f(\gamma_t), \gamma_t - \Theta_t(x) \rangle + \langle \dot{\gamma}_t + \nabla f(\gamma_t), \gamma_t - \Theta_t(x) \rangle \end{aligned}$$

$$\begin{aligned}
&\leq \beta \|\gamma_t - \Theta_t(x)\|^2 + \frac{1}{2} \|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2 + \frac{1}{2} \|\gamma_t - \Theta_t(x)\|^2 \\
&= \left(\beta + \frac{1}{2}\right) g_t + \frac{1}{2} \|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2,
\end{aligned} \tag{B.3.97}$$

where in the last inequality we used the smoothness of f (Assumption B.1) and Young's inequality.

Rewriting (B.3.97), we have

$$\begin{aligned}
\frac{d}{dt} \left(e^{-(\beta+\frac{1}{2})t} g_t \right) &\leq e^{-(\beta+1)t} \frac{1}{2} \|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2 \\
&\leq \frac{1}{2} \|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2,
\end{aligned} \tag{B.3.98}$$

so that integrating yields, for any $t \in [0, T]$,

$$g_t \leq \frac{e^{(\beta+\frac{1}{2})t}}{2} \int_0^t \|\dot{\gamma}_s + \nabla f(\gamma_s)\|^2 ds \tag{B.3.99}$$

where we used that $\Theta_0(x) = x = \gamma_0$.

Furthermore, introducing a , we have

$$\begin{aligned}
g_t &\leq a e^{(\beta+\frac{1}{2})t} \int_0^t \frac{\|\dot{\gamma}_s + \nabla f(\gamma_s)\|^2}{\sigma_\infty^2(f(\gamma_s))} ds \\
&\leq a e^{(\beta+\frac{1}{2})t} \mathcal{S}_{0,t}(\gamma),
\end{aligned} \tag{B.3.100}$$

by Lemma B.1.2.

Hence, we have that, for any $t \in [0, T]$,

$$\|\gamma_t - \Theta_t(x)\|^2 \leq 2a e^{(\beta+\frac{1}{2})t} \mathcal{S}_{0,t}(\gamma). \tag{B.3.101}$$

Therefore, if we consider γ such that, in addition, $\mathcal{S}_T(\gamma) \leq \frac{e^{-(\beta+\frac{1}{2})S}}{2a} \times \frac{\delta'^2}{4}$, we have that, for $t = \min(T, S)$,

$$\|\gamma_t - \Theta_t(x)\| \leq \frac{\delta'}{2}. \tag{B.3.102}$$

We now distinguish two cases:

- If $t = S$ i.e., $S \leq T$, then (B.3.102) combined with (B.3.95) yields that $\gamma_S \in \mathcal{U}_{\delta'}(C)$. But γ_T must be in $\mathbb{R}^d \setminus \text{Attr}(C)$ and so in $\mathbb{R}^d \setminus \mathcal{U}_\delta(C)$, so that one must have

$$\mathcal{S}_T(\gamma) \geq \mathcal{B}(\mathcal{U}_{\delta'}(C), \mathbb{R}^d \setminus \text{Attr}(C)). \tag{B.3.103}$$

- If $t = T$ i.e., $T \leq S$, we extend the path γ to $\varphi \in \mathcal{C}([0, S])$ such that $\varphi_t = \gamma_t$ for $t \in [0, T]$ and $\varphi_t = \Theta_t(x)$ for $t \in [T, S]$. In particular, $\mathcal{S}_S(\varphi) = \mathcal{S}_T(\gamma) \leq \frac{e^{-(\beta+\frac{1}{2})T}}{2a} \times \frac{\delta'^2}{4}$ so that the same computations leading to (B.3.102) can be applied to φ to yield that

$$\|\varphi_S - \Theta_S(x)\| \leq \frac{\delta'}{2}. \tag{B.3.104}$$

This means that $\varphi_S \in \mathcal{U}_{\delta'}(C)$ which is included in $\text{Attr}(C)$. This is contradiction since this means that γ_T is in $\text{Attr}(C)$. Therefore, this case cannot happen.

Therefore, we have shown that, for any $\gamma \in \mathcal{C}([0, T])$ such that $\gamma_0 = x$ and $\gamma_T \in \mathbb{R}^d \setminus \text{Attr}(C)$, it holds that

$$S_T(\gamma) \geq \min \left(1, \frac{e^{-(\beta+\frac{1}{2})T}}{2a} \times \frac{\delta'^2}{4}, \mathcal{B}(\mathcal{U}_{\delta'}(C), \mathbb{R}^d \setminus \mathcal{U}_{\delta}(C)) \right), \quad (\text{B.3.105})$$

which is positive by (B.3.93). $\blacksquare$

We are now ready to prove the main result of this section.

Theorem B.19. *Consider $x_0 \in \mathbb{R}^d$ that belongs to $\text{Attr}(C_i)$ for some $i \in \{1, \dots, N_{\text{crit}}\}$ such that C_i is minimizing. Then, for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $C_1, \dots, C_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$, an event H and a positive constant A such that for all $0 < \eta < \eta_0$,*

$$\mathbb{P}_{x_0}(H) \geq 1 - e^{-A/\eta} \quad (\text{B.3.106})$$

and,

$$e^{\frac{E(\mathcal{Q}|i)-\varepsilon}{\eta}} \leq \mathbb{E}_{x_0}[\tau_{\mathcal{Q}}^\eta | H] \leq e^{\frac{E(\mathcal{Q}|i)+\varepsilon}{\eta}}. \quad (\text{B.3.107})$$

Proof. First, Lemma B.3.18 ensures that $\mathcal{B}(x_0, \mathbb{R}^d \setminus \text{Attr}(C_i)) > 0$ so that, for any $j \in \mathcal{V}$, $j \neq i$, it holds $\mathcal{B}(x_0, C_j) > 0$. Define

$$A := \min_{j \in \mathcal{V}, j \neq i} \mathcal{B}(x_0, C_j) > 0. \quad (\text{B.3.108})$$

Let us now apply Lemma B.2.11 to obtain that, for any small enough neighborhoods $\mathcal{C}_i^\delta$ of C_i , $i = 1, \dots, N_{\text{crit}}$, there is some $\eta_0 > 0$ such that for all $j \in \{1, \dots, N_{\text{crit}}\}$, $0 < \eta < \eta_0$,

$$\mathcal{Q}C^\delta(x, \mathcal{C}_j^\delta) \leq e^{-\frac{A}{2\eta}}. \quad (\text{B.3.109})$$

Now define the event H as

$$H := \left\{ x_{\tau_1^\eta}^\eta \in \mathcal{C}_i^\delta \right\} \quad (\text{B.3.110})$$

with the notation of Definition B.12. (B.3.109) then ensures that

$$\mathbb{P}_{x_0}(H) \geq 1 - e^{-\frac{A}{4\eta}}, \quad (\text{B.3.111})$$

provided that η is small enough so that $e^{-\frac{A}{4\eta}} N_{\text{crit}} \leq 1$. If $i \in \mathcal{Q}$, the result is immediate. In the following, we assume that $i \notin \mathcal{Q}$.

Let us now invoke Lemma B.3.6 to obtain that, at the potential cost of requiring smaller $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $C_1, \dots, C_{N_{\text{crit}}}$ or of reducing η_0 , it holds that, for any $0 < \eta < \eta_0$, $x \in x_0$,

$$e^{\frac{E(\mathcal{Q}|i)-\varepsilon}{\eta}} \leq \mathbb{E}_x[\tau_{\mathcal{Q}}^\eta] \leq e^{\frac{E(\mathcal{Q}|i)+\varepsilon}{\eta}}. \quad (\text{B.3.112})$$

To conclude the proof, write, by the strong Markov property,

$$\begin{aligned} \mathbb{E}_{x_0}[\tau_{\mathcal{Q}}^\eta | H] &= \frac{\mathbb{E}_{x_0}[\mathbb{1}_H \tau_{\mathcal{Q}}^\eta]}{\mathbb{P}_{x_0}(H)} \\ &= \frac{\mathbb{E}_{x_0} \left[\mathbb{1}_H \left(1 + \mathbb{E}_{x_{\tau_1^\eta}^\eta}[\tau_{\mathcal{Q}}^\eta] \right) \right]}{\mathbb{P}_{x_0}(H)} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{\mathbb{E}_{x_0}[\mathbf{1}_H] e^{\frac{E(\mathcal{Q}|i)+2\varepsilon}{\eta}}}{\mathbb{P}_{x_0}(H)} \\
&= e^{\frac{E(\mathcal{Q}|i)+2\varepsilon}{\eta}}, \tag{B.3.113}
\end{aligned}$$

where we used (B.3.112) in the last inequality and taking η small enough so that $e^{\frac{\varepsilon}{\eta}} \geq 2$. We have thus shown the RHS of the inequality of the statement, and the LHS is obtained by the same argument. $\blacksquare$

Similarly to how Theorem B.8 was obtained from Theorem B.5, we immediately obtain the following corollary.

Theorem B.20. Consider $x_0 \in \mathbb{R}^d$ that belongs to $\text{Attr}(\mathcal{C}_i)$ for some $i \in \{1, \dots, N_{\text{crit}}\}$ such that $\mathcal{C}_i$ is minimizing. Then, for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$, an event H and a positive constant A such that for all $0 < \eta < \eta_0$,

$$\mathbb{P}_{x_0}(H) \geq 1 - e^{-A/\eta} \tag{B.3.114}$$

and,

$$\mathbb{E}_{x_0}[\tau | H] \leq e^{\frac{E(\mathcal{Q}|i)+\varepsilon}{\eta}}. \tag{B.3.115}$$

We now turn our attention to the lower-bound counterpart of Theorem B.20.

Theorem B.21. Consider $x_0 \in \mathbb{R}^d$ that belongs to $\text{Attr}(\mathcal{C}_i)$ for some $i \in \{1, \dots, N_{\text{crit}}\}$ such that $\mathcal{C}_i$ is minimizing. Under Assumption B.20 and assuming that,

$$\frac{\mu R^2}{1152\bar{\sigma}_\infty^2} > \mathcal{B}_\infty - E(\mathcal{Q}|i), \tag{B.3.116}$$

for any $\varepsilon > 0$, for $\mathcal{C}_1^\delta, \dots, \mathcal{C}_{N_{\text{crit}}}^\delta$ neighborhoods of $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{crit}}}$ small enough, there is $\eta_0 > 0$, an event H and a positive constant A such that for all $0 < \eta < \eta_0$,

$$\mathbb{P}_{x_0}(H) \geq 1 - e^{-A/\eta} \tag{B.3.117}$$

and,

$$\mathbb{E}_{x_0}[\tau] \geq e^{\frac{E(\mathcal{Q}|i)-\varepsilon}{\eta}}. \tag{B.3.118}$$

Theorem B.21 is obtained by combining Lemma B.3.13 with Lemma B.3.18 in the same way as Theorem B.19 was obtained from Lemma B.3.6.

B.3.5 Full transition costs

Definition B.22. We define, for $i, j \in \{1, \dots, N_{\text{crit}}\}$,

$$\begin{aligned}
C_{i,j} &:= \inf\{\mathcal{S}_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 \in \mathcal{C}_i, \xi_{N-1} \in \mathcal{C}_j\} \\
&= \inf\{\mathcal{S}_T(\gamma) : T \in \mathbb{N}, \gamma \in \mathcal{D}_r(T), \gamma_0 \in \mathcal{C}_i, \gamma_T \in \mathcal{C}_j\}. \tag{B.3.119}
\end{aligned}$$

With the same arguments as Lemma B.2.2, we can show the alternate expression for $C_{i,j}$.

Lemma B.3.22. For any $i, j \in \{1, \dots, N_{\text{crit}}\}$, it holds that

$$C_{i,j} = \lim_{\delta \rightarrow 0} C_{i,j}^\delta, \tag{B.3.120}$$

where

$$\begin{aligned} C_{i,j}^\delta &:= \inf\{\mathcal{S}_N(\xi) : N \geq 1, \xi \in \mathcal{D}_r(N), \xi_0 \in \mathcal{U}_\delta(\mathcal{C}_i), \xi_{N-1} \in \mathcal{U}_\delta(\mathcal{C}_j)\} \\ &= \inf\{\mathcal{S}_T(\gamma) : T \in \mathbb{N}, \gamma \in \mathcal{D}_r(T), \gamma_0 \in \mathcal{U}_\delta(\mathcal{C}_i), \gamma_T \in \mathcal{U}_\delta(\mathcal{C}_j)\}. \end{aligned} \quad (\text{B.3.121})$$

We can also relate $C_{i,j}$ in terms $B_{i,j}$.

Lemma B.3.23. *For any $i \in \{1, \dots, K\}$, $j \in \{1, \dots, N_{\text{crit}}\}$, it holds that*

$$C_{i,j} \leq B_{i,j}, \quad (\text{B.3.122})$$

and, if $C_{i,j} < B_{i,j}$, then there must exist $l \in \mathcal{Q}$ such that $B_{i,l} \leq C_{i,j} < B_{i,j}$.

This lemma is a direct consequence of the definitions of $C_{i,j}$ (Definition B.22) and $B_{i,j}$ (Definition B.14) and is therefore omitted.

Lemma B.3.24. *For any $i \in \{1, \dots, N_{\text{crit}}\}$,*

$$E(\mathcal{Q} | i) \leq \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\} - \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(i \not\rightarrow \mathcal{Q}) \right\}. \quad (\text{B.3.123})$$

Proof. By Definition B.18, we have that

$$E(\mathcal{Q} | i) = E(\mathcal{Q}) - E(i \not\rightarrow \mathcal{Q}). \quad (\text{B.3.124})$$

Let us begin by lower-bounding the second term. By Lemma B.3.23, it holds that

$$\begin{aligned} E(i \not\rightarrow \mathcal{Q}) &= \min \left\{ \sum_{k \rightarrow l \in g} B_{k,l} : g \in \mathcal{G}(i \not\rightarrow \mathcal{Q}) \right\} \\ &\geq \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(i \not\rightarrow \mathcal{Q}) \right\}. \end{aligned} \quad (\text{B.3.125})$$

Now, let us turn our attention to the first term: by Lemma B.3.23, we have that

$$\begin{aligned} E(\mathcal{Q}) &:= \min \left\{ \sum_{k \rightarrow l \in g} B_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\} \\ &\geq \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\}. \end{aligned} \quad (\text{B.3.126})$$

But, for any $g \in \mathcal{G}(\mathcal{Q})$ which reaches that minimum, all its edges $k \rightarrow l$ must be such $B_{k,l} = C_{k,l}$ by Lemma B.3.23 (otherwise we would replace it by an edge to $\mathcal{Q}$ with a lower cost). Therefore, we must also have that

$$\min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\} \geq \min \left\{ \sum_{k \rightarrow l \in g} B_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\}, \quad (\text{B.3.127})$$

which shows that

$$E(\mathcal{Q}) = \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\}. \quad (\text{B.3.128})$$

This concludes the proof. $\blacksquare$

Finally, let us restate a result from (Azizian et al., 2024b) that will be relevant in the following.

Lemma B.3.25 ((Azizian et al., 2024b, Lem. D.31)). *If $\mathcal{C}_i$ is not asymptotically stable, then there exists $j \in \mathcal{V}$ such that $B_{i,j} = 0$ and such that $\mathcal{C}_j$ is asymptotically stable.*

B.3.6 Link to topological properties of the loss landscape

The goal of this section is to prove the following result.

Proposition B.3.26. *The following properties are equivalent:*

- (i) For all $x \in \mathbb{R}^d$, $E(\mathcal{Q} | x) = 0$.
- (ii) For all $i \in \{1, \dots, K\}$, $E(\mathcal{Q} | i) = 0$.
- (iii) For all $i \in \{1, \dots, K\}$, $\mathcal{K}_i$ is not locally minimizing.

Proof. Items (i) and (ii) are equivalent by definition of $E(\mathcal{Q} | x)$ (Definition B.19). We thus focus on the equivalence of (ii) and (iii).

First, let us consider the case where there is some $i \in \{1, \dots, K\}$ such that $\mathcal{K}_i$ is locally minimizing. By Lemma B.3.17, there exists $\delta > 0$ small enough such that

$$\mathcal{B}(\mathcal{K}_i, \mathbb{R}^d \setminus \mathcal{U}_\delta(\mathcal{K}_i)) > 0. \quad (\text{B.3.129})$$

As a consequence, for any $j \neq i$, $B_{i,j} > 0$.

Consider $g \in \mathcal{G}(\mathcal{V})$ such that

$$E(\mathcal{Q}) = \sum_{k \rightarrow l \in g} B_{k,l}. \quad (\text{B.3.130})$$

Consider the graph g' obtained by removing the edge in graph that exits i . This edge must have positive cost and therefore

$$\sum_{k \rightarrow l \in g} B_{k,l} > \sum_{k \rightarrow l \in g'} B_{k,l} \geq E(i \not\rightarrow \mathcal{Q}). \quad (\text{B.3.131})$$

Combining (B.3.130) and (B.3.131), we obtain that $E(\mathcal{Q} | i) > 0$.

Let us now turn to the case where there is no $i \in \{1, \dots, K\}$ such that $\mathcal{K}_i$ is locally minimizing. Take any $i \in \{1, \dots, K\}$. We bound $E(\mathcal{Q} | i)$ with Lemma B.3.24:

$$E(\mathcal{Q} | i) \leq \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\} - \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(i \not\rightarrow \mathcal{Q}) \right\}. \quad (\text{B.3.132})$$

Take any $g \in \mathcal{G}(i \not\rightarrow \mathcal{Q})$. By Lemma B.3.25, there exists $j \in \mathcal{V}$ such that $B_{i,j} = 0$ and $\mathcal{K}_j$ is asymptotically stable. By assumption, j must thus belong to $\mathcal{Q}$. Consider $g' \in \mathcal{G}(\mathcal{Q})$ obtained by adding the edge $i \rightarrow j$ to g . (B.3.132) then yields that $E(\mathcal{Q} | i) = 0$. $\blacksquare$

B.4 POTENTIAL-BASED BOUNDS

B.4.1 Noise assumptions

Assumption B.23 introduces key requirements for the noise behavior through bounds on $\bar{\mathcal{H}}$ and $\bar{\mathcal{L}}$ around the gradient $\nabla f(x)$. Intuitively, the bound on $\bar{\mathcal{H}}$ controls how large the noise can be (upper bound), while the bound on $\bar{\mathcal{L}}$ ensures the noise maintains sufficient variance in all directions (lower bound) around the gradient. We then present two ways to satisfy these requirements:

GAUSSIAN NOISE. The first approach, detailed in [Lemma B.4.2](#), considers the case when the noise follows a truncated Gaussian distribution. This is a standard Gaussian with variance $\sigma^2(f(x))$, but restricted to stay within a ball of radius $R(f(x))$.

SUPPORT AND COVARIANCE CONDITIONS. [Lemma B.4.1](#) provides a more general approach with two main requirements:

- A support condition ensuring the noise can be large enough in any direction around $\nabla f(x)$ with some positive probability
- A lower bound on the covariance matrix of the noise

This second approach is more flexible as it allows for non-Gaussian noise distributions. A key example is discrete noise distributions, where $U(x, \omega)$ takes values in a finite set. The support condition then requires that some points in the support are sufficiently far from $\nabla f(x)$ in every direction.

Assumption B.23. For some $X \subset \mathbb{R}^d$, there exist $\bar{\sigma}^2, \underline{\sigma}^2 : \mathbb{R} \rightarrow (0, +\infty)$ continuous functions such that, for all $x \in X$, it holds that

$$\begin{aligned} \forall p \in \mathbb{R}^d \text{ s.t. } \|p\| \leq \frac{2\|\nabla f(x)\|}{\bar{\sigma}^2(f(x))}, \quad \bar{\mathcal{H}}(x, p) &\leq \frac{\bar{\sigma}^2(f(x))\|p\|^2}{2} \\ \forall v \in \mathbb{R}^d \text{ s.t. } \|v - \nabla f(x)\| \leq \|\nabla f(x)\|, \quad \bar{\mathcal{L}}(x, v) &\leq \frac{\|v\|^2}{2\underline{\sigma}^2(f(x))}. \end{aligned} \quad (\text{B.4.1})$$

Remark 8. [Assumption B.23](#) is a generalized and detailed version of [Assumption B.23](#). Indeed, our blanket assumptions on the noise [Assumption B.3](#) already imply an upper-bound on $\bar{\mathcal{H}}(x, p)$ with $\bar{\sigma}^2 = \sigma_\infty^2$ for any $x, p \in \mathbb{R}^d$. [Assumption B.23](#) allows for refined results when a tighter upper-bound on $\bar{\mathcal{H}}(x, p)$ is available for a restricted set of x and p .

Note that, in general, we will have $\bar{\sigma}^2 \geq \underline{\sigma}^2$, hence the notation.

Lemma B.4.1. Assume that, for some $X \subset \mathbb{R}^d$ relatively compact,

- there exists $a > 0$ such that,

$$b := \inf\{\mathbb{P}(\langle U(x, \omega) - \nabla f(x), q \rangle \geq a + \|\nabla f(x)\|) : x \in X, q \in \mathbb{S}^{d-1}\} > 0, \quad (\text{B.4.2})$$

- there exists $\zeta^2 : \mathbb{R} \rightarrow (0, +\infty)$ continuous function such that, for all $x \in X$, it holds that

$$\text{cov } U(x, \omega) \geq \zeta^2(f(x))I_d, \quad (\text{B.4.3})$$

Then, for any $x \in X, v \in \mathbb{R}^d$ such that $\|v - \nabla f(x)\| \leq \|\nabla f(x)\|$, it holds that

$$\bar{\mathcal{L}}(x, v) \leq \max\left(1, \frac{\log b^{-1}}{ac}\right) \frac{\|v\|^2}{\zeta^2(f(x))}, \quad (\text{B.4.4})$$

where $c > 0$ is a constant small enough such that

$$\forall x \in X, p \in \overline{\mathbb{B}}(0, c), \quad \|\text{Hess}_p \bar{\mathcal{H}}(x, p) - \text{cov } U(x, \omega)\| \leq \frac{1}{2} \zeta^2(f(x)), \quad (\text{B.4.5})$$

for the matrix norm associated with the Euclidean norm.

Note that c always exists by the relative compactness of X and the fact that

$$\text{Hess}_p \bar{\mathcal{H}}(x, 0) = \text{cov } U(x, \omega). \quad (\text{B.4.6})$$

Proof. The first part of the proof consists in showing that, for $x \in X, p \in \mathbb{R}^d$, it holds that

$$\bar{\mathcal{H}}(x, p) \leq \frac{\zeta^2(f(x)) \min(c, \|p\|) \|p\|}{4}. \quad (\text{B.4.7})$$

Let us first consider the case where $\|p\| \leq c$. Since both $\bar{\mathcal{H}}(x, 0) = 0$ and $\nabla_p \bar{\mathcal{H}}(x, 0) = \mathbb{E}[U(x, \omega)] = 0$, we have that

$$\begin{aligned} \bar{\mathcal{H}}(x, p) &= \frac{1}{2} \int_0^1 \langle p, \text{Hess}_p \bar{\mathcal{H}}(x, tp) p \rangle dt \\ &\geq \frac{1}{2} \left(\int_0^1 \langle p, \text{cov } U(x, \omega) p \rangle dt - \frac{1}{2} \zeta^2(f(x)) \|p\|^2 \right) \\ &\geq \frac{\zeta^2(f(x)) \|p\|^2}{4}, \end{aligned} \quad (\text{B.4.8})$$

where we successively used Cauchy-Schwarz inequality, (B.4.5) and (B.4.3). (B.4.8) readily yields (B.4.7) when $\|p\| \leq c$.

Let us now consider the case where $\|p\| > c$. By concavity of the function $s \mapsto s^{c/\|p\|} s$ on $\mathbb{R}_+$ and Jensen's inequality, we have that

$$\begin{aligned} \bar{\mathcal{H}}(x, p) &\geq \frac{\|p\|}{c} \bar{\mathcal{H}}\left(x, c \frac{p}{\|p\|}\right) \\ &\geq \frac{\|p\|}{c} \frac{\zeta^2(f(x)) \|c \frac{p}{\|p\|}\|^2}{4} = \frac{\zeta^2(f(x)) c \|p\|}{4}, \end{aligned} \quad (\text{B.4.9})$$

where we used (B.4.8) in the last inequality. This concludes the proof of (B.4.7).

The second step of this proof now consists in showing that, for $x \in X, v \in \mathbb{R}^d$ such that $\|v - \nabla f(x)\| \leq \|\nabla f(x)\|$, it holds that

$$\arg \max_{p \in \mathbb{R}^d} \{\langle p, v \rangle - \bar{\mathcal{H}}(x, p)\} = \arg \min_{p \in \mathbb{R}^d} \{\bar{\mathcal{H}}(x, p) - \langle p, v \rangle\} \subset \overline{\mathbb{B}}\left(0, \frac{\log b^{-1}}{a}\right). \quad (\text{B.4.10})$$

For $x \in X, p \in \mathbb{R}^d$, we can lower-bound $\bar{\mathcal{H}}(x, p)$ using (B.4.2):

$$\begin{aligned} \bar{\mathcal{H}}(x, p) &= \log \mathbb{E}[\exp(\langle p, U(x, \omega) \rangle)] \\ &= \langle p, \nabla f(x) \rangle + \log \mathbb{E}[\exp(\langle p, U(x, \omega) - \nabla f(x) \rangle)] \\ &\geq \langle p, \nabla f(x) \rangle + \log \mathbb{E}\left[e^{\|p\|(a + \|\nabla f(x)\|)} \mathbb{1}\left\{\left\langle U(x, \omega) - \nabla f(x), \frac{p}{\|p\|} \right\rangle \geq a + \|\nabla f(x)\|\right\}\right] \\ &\geq \langle p, \nabla f(x) \rangle + \|p\|(a + \|\nabla f(x)\|) + \log b. \end{aligned} \quad (\text{B.4.11})$$

As a consequence, for $v \in \mathbb{R}^d$ such that $\|v - \nabla f(x)\| \leq \|\nabla f(x)\|$, we obtain that

$$\bar{\mathcal{H}}(x, p) - \langle p, v \rangle \geq \langle p, \nabla f(x) - v \rangle + \|p\|(a + \|\nabla f(x)\|) + \log b$$

$$\geq \|p\|a + \log b, \quad (\text{B.4.12})$$

where we used Cauchy-Schwarz inequality and the condition on v .

In particular, (B.4.12) implies that $p \mapsto \tilde{\mathcal{H}}(x, p) - \langle p, v \rangle$ is coercive on $\mathbb{R}^d$ and therefore $\arg \min_{p \in \mathbb{R}^d} \{\tilde{\mathcal{H}}(x, p) - \langle p, v \rangle\}$ is well-defined.

Remarking that $\tilde{\mathcal{H}}(x, 0) - \langle 0, v \rangle = 0$, we can now conclude that $\arg \min_{p \in \mathbb{R}^d} \{\tilde{\mathcal{H}}(x, p) - \langle p, v \rangle\}$ must be included in

$$\{p \in \mathbb{R}^d : \|p\|a + \log b \leq 0\} = \overline{\mathbb{B}}\left(0, \frac{\log b^{-1}}{a}\right), \quad (\text{B.4.13})$$

which completes the proof of (B.4.10).

To conclude the proof of this lemma, we now use (B.4.10) to obtain that, for $x \in X, v \in \mathbb{R}^d$ such that $\|v - \nabla f(x)\| \leq \|\nabla f(x)\|$,

$$\begin{aligned} \tilde{\mathcal{L}}(x, v) &= \sup_{p \in \mathbb{R}^d} \{\langle p, v \rangle - \tilde{\mathcal{H}}(x, p)\} \\ &= \sup \left\{ \langle p, v \rangle - \tilde{\mathcal{H}}(x, p) : p \in \overline{\mathbb{B}}\left(0, \frac{\log b^{-1}}{a}\right) \right\} \\ &\leq \sup \left\{ \langle p, v \rangle - \frac{\zeta^2(f(x)) \min(c, \|p\|) \|p\|}{4} : p \in \overline{\mathbb{B}}\left(0, \frac{\log b^{-1}}{a}\right) \right\}, \end{aligned} \quad (\text{B.4.14})$$

where we used (B.4.7) in the last inequality.

Now, we note that, for any $p \in \overline{\mathbb{B}}\left(0, \frac{\log b^{-1}}{a}\right)$, it holds that

$$\begin{aligned} \frac{\zeta^2(f(x)) \min(c, \|p\|) \|p\|}{4} &\geq \begin{cases} \frac{\zeta^2(f(x)) \|p\|^2}{4} & \text{if } \|p\| \leq c \\ \frac{a}{\log b^{-1}} \frac{\zeta^2(f(x)) c \|p\|^2}{4} & \text{if } \|p\| > c \end{cases} \\ &\geq \min\left(1, \frac{ac}{\log b^{-1}}\right) \frac{\zeta^2(f(x)) \|p\|^2}{4}. \end{aligned} \quad (\text{B.4.15})$$

Plugging (B.4.15) into (B.4.14), we obtain that

$$\begin{aligned} \tilde{\mathcal{L}}(x, v) &\leq \sup \left\{ \langle p, v \rangle - \min\left(1, \frac{ac}{\log b^{-1}}\right) \frac{\zeta^2(f(x)) \|p\|^2}{4} : p \in \overline{\mathbb{B}}\left(0, \frac{\log b^{-1}}{a}\right) \right\} \\ &\leq \sup \left\{ \langle p, v \rangle - \min\left(1, \frac{ac}{\log b^{-1}}\right) \frac{\zeta^2(f(x)) \|p\|^2}{4} : p \in \mathbb{R}^d \right\} \\ &= \max\left(1, \frac{\log b^{-1}}{ac}\right) \frac{\|v\|^2}{\zeta^2(f(x))}, \end{aligned} \quad (\text{B.4.16})$$

which concludes the proof of this lemma. $\blacksquare$

The following lemma is an immediate consequence of Lemma B.5.1.

Lemma B.4.2. Consider $\varepsilon > 0$. Assume that $U(x, \omega)$ follows a Gaussian distribution $\mathcal{N}(0, \sigma^2(f(x))I_d)$ conditioned on being in $\overline{\mathbb{B}}(0, R(f(x)))$ for all $x \in \mathbb{R}^d$ and some continuous function $\sigma^2, R : \mathbb{R} \rightarrow (0, +\infty)$. Assume that $8\|\nabla f(x)\| \leq R(f(x))$ for all $x \in \mathbb{R}^d$ and that $\inf_{x \in \mathbb{R}^d} \frac{R(f(x))}{\sigma(f(x))} \geq \mathcal{O}(\log 1/\varepsilon)$. Then Assumption B.23 holds with $X = \mathbb{R}^d$, $\bar{\sigma}^2(f(x)) = (1 + \varepsilon)\sigma^2(f(x))$ and $\underline{\sigma}^2(f(x)) = (1 - \varepsilon)\sigma^2(f(x))$ for all $x \in \mathbb{R}^d$.

B.4.2 Potentials and path reversal

Let us now define the potentials associated to the variance functions $\bar{\sigma}^2$ and $\underline{\sigma}^2$ introduced in [Assumption B.23](#).

Definition B.24. Given $\bar{\sigma}^2, \underline{\sigma}^2 : \mathbb{R} \rightarrow (0, +\infty)$ continuous functions, we define the potentials $\bar{V}, \underline{V} : \mathbb{R} \rightarrow \mathbb{R}$ as

$$\bar{V}(x) := 2\bar{\alpha}(f(x)) \quad \text{where } \bar{\alpha} : \mathbb{R} \rightarrow \mathbb{R} \text{ is a primitive of } 1/\bar{\sigma}^2 \quad (\text{B.4.17})$$

$$\underline{V}(x) := 2\underline{\alpha}(f(x)) \quad \text{where } \underline{\alpha} : \mathbb{R} \rightarrow \mathbb{R} \text{ is a primitive of } 1/\underline{\sigma}^2. \quad (\text{B.4.18})$$

We begin by adapting ([Azizian et al., 2024b](#), Lem .E.1) to the current setting.

Lemma B.4.3. Consider $\gamma \in \mathcal{C}([0, T])$. Then, there exists $\tilde{\gamma} \in \mathcal{C}([0, S])$ a reparametrization of γ such that, for any $t \in [0, S]$,

$$\|\dot{\tilde{\gamma}}_s\| = \|\nabla f(\tilde{\gamma}_s)\|. \quad (\text{B.4.19})$$

and, under [Assumption B.23](#),

$$\mathcal{S}_T(\gamma) \geq \int_0^S \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2\bar{\sigma}^2(f(\tilde{\gamma}_s))} ds. \quad (\text{B.4.20})$$

Proof. By the proof [Freidlin and Wentzell \(1998b, Chap. 4, Lem. 3.1\)](#), there exists $t(s)$ change of time such that, with $\tilde{\gamma}_s = \gamma_{t(s)}$, $\|\dot{\tilde{\gamma}}_s\| = \|\nabla f(\tilde{\gamma}_s)\|$.

We have that

$$\mathcal{S}_T(\gamma) = \int_0^{t^{-1}(T)} \dot{t}(s) \mathcal{L}(\tilde{\gamma}_s, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s) ds, \quad (\text{B.4.21})$$

so it suffices to bound $\mathcal{L}(\tilde{\gamma}_s, \dot{\tilde{\gamma}}_s)$ from below: by definition, we have

$$\begin{aligned} \mathcal{L}(\tilde{\gamma}_s, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s) &\geq \sup \left\{ \langle p, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s) \rangle - \bar{\mathcal{H}}(\tilde{\gamma}_s, p) : \|p\| \leq \frac{2\|\nabla f(\tilde{\gamma}_s)\|}{\bar{\sigma}^2(f(\tilde{\gamma}_s))} \right\} \\ &\geq \sup \left\{ \langle p, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s) \rangle - \frac{\bar{\sigma}^2(f(\tilde{\gamma}_s))}{2} \|p\|^2 : \|p\| \leq \frac{2\|\nabla f(\tilde{\gamma}_s)\|}{\bar{\sigma}^2(f(\tilde{\gamma}_s))} \right\}, \end{aligned} \quad (\text{B.4.22})$$

by [Assumption B.23](#). Applying [Lemma B.5.2](#) with $v \leftarrow (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s$, $w \leftarrow \nabla f(\tilde{\gamma}_s)$ and $\lambda \leftarrow (\dot{t}(s))^{-1}$, now exactly yields, for almost all s ,

$$\begin{aligned} \mathcal{L}(\tilde{\gamma}_s, (\dot{t}(s))^{-1} \dot{\tilde{\gamma}}_s) &\geq (\dot{t}(s))^{-1} \sup \left\{ \langle p, \dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s) \rangle - \frac{\bar{\sigma}^2(f(\tilde{\gamma}_s))}{2} \|p\|^2 : \|p\| \leq \frac{2\|\nabla f(\tilde{\gamma}_s)\|}{\bar{\sigma}^2(f(\tilde{\gamma}_s))} \right\} \\ &= (\dot{t}(s))^{-1} \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2\bar{\sigma}^2(f(\tilde{\gamma}_s))}, \end{aligned} \quad (\text{B.4.23})$$

since $p \mapsto \langle p, \dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s) \rangle - \frac{\bar{\sigma}^2(f(\tilde{\gamma}_s))}{2} \|p\|^2$ reaches its maximum at $p^* = \frac{\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)}{\bar{\sigma}^2(f(\tilde{\gamma}_s))}$ whose norm satisfies $\|p^*\| \leq \frac{2\|\nabla f(\tilde{\gamma}_s)\|}{\bar{\sigma}^2(f(\tilde{\gamma}_s))}$. $\blacksquare$

An important consequence of [Assumption B.23](#) is that the action cost of going from x to x' can be related to the action cost of going from x' to x . This is formalized in [Lemma B.4.4](#).

Lemma B.4.4. *For any $T > 0$, $x, x' \in \mathbb{R}^d$, $\gamma \in \mathcal{C}([0, T])$ such that $\gamma_0 = x$ and $\gamma_T = x'$, under [Assumption B.23](#), if γ is contained in X , it holds that,*

$$\mathcal{S}_T(\gamma) \geq \inf_{x \in X} \frac{\sigma^2(f(x))}{\bar{\sigma}^2(f(x))} \times \left(\mathcal{B}(x, x') + \bar{V}(x') - \bar{V}(x) \right), \quad (\text{B.4.24})$$

Proof. Let us first invoke [Lemma B.4.3](#): there exists $\tilde{\gamma} \in \mathcal{C}([0, S])$ a reparametrization of γ such that, for any $t \in [0, S]$,

$$\|\dot{\tilde{\gamma}}_s\| = \|\nabla f(\tilde{\gamma}_s)\|. \quad (\text{B.4.25})$$

and,

$$\begin{aligned} \mathcal{S}_T(\gamma) &\geq \int_0^S \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2\bar{\sigma}^2(f(\tilde{\gamma}_s))} ds \\ &\geq \inf_{x \in X} \frac{\sigma^2(f(x))}{\bar{\sigma}^2(f(x))} \times \left(\int_0^S \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2\bar{\sigma}^2(f(\tilde{\gamma}_s))} ds \right). \end{aligned} \quad (\text{B.4.26})$$

It now suffices to lower-bound the integral on the RHS of (B.4.26). We have

$$\int_0^S \frac{\|\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2\bar{\sigma}^2(f(\tilde{\gamma}_s))} ds = \int_0^S \frac{\|-\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2\bar{\sigma}^2(f(\tilde{\gamma}_s))} ds + \int_0^S \frac{2\langle \dot{\tilde{\gamma}}_s, \nabla f(\tilde{\gamma}_s) \rangle}{\bar{\sigma}^2(f(\tilde{\gamma}_s))} ds. \quad (\text{B.4.27})$$

By definition of $\bar{V}$ ([Definition B.24](#)) the second integral is equal to

$$\begin{aligned} \int_0^S \frac{2\langle \dot{\tilde{\gamma}}_s, \nabla f(\tilde{\gamma}_s) \rangle}{\bar{\sigma}^2(f(\tilde{\gamma}_s))} ds &= \int_0^S \langle \dot{\tilde{\gamma}}_s, \nabla \bar{V}(\tilde{\gamma}_s) \rangle ds \\ &= \bar{V}(x') - \bar{V}(x). \end{aligned} \quad (\text{B.4.28})$$

For the first integral in (B.4.27), define the path $(\varphi_s)_{s \in [0, S]}$ by $\varphi_s = \tilde{\gamma}_{S-s}$. We have that

$$\begin{aligned} \int_0^S \frac{\|-\dot{\tilde{\gamma}}_s + \nabla f(\tilde{\gamma}_s)\|^2}{2\bar{\sigma}^2(f(\tilde{\gamma}_s))} ds &= \int_0^S \frac{\|\dot{\varphi}_s + \nabla f(\varphi_s)\|^2}{2\bar{\sigma}^2(f(\varphi_s))} ds \\ &\geq \int_0^S \tilde{\mathcal{L}}(\varphi_s, \dot{\varphi}_s + \nabla f(\varphi_s)) ds \\ &= \mathcal{S}_S(\varphi) \\ &\geq \mathcal{B}(x, x'), \end{aligned} \quad (\text{B.4.29})$$

where we used [Assumption B.23](#) in the first inequality with $v \leftarrow \dot{\varphi}_s + \nabla f(\varphi_s)$ whose norm satisfies the condition of [Assumption B.23](#) by (B.4.25). Plugging (B.4.28) and (B.4.29) into (B.4.27) and then into (B.4.26) yields the desired result. $\blacksquare$

B.4.3 Graph representation

Let us reuse the notation of [Appendix B.3](#). In [Appendix B.3](#), we consider graphs over the set of vertices $\mathcal{V} = \{1, \dots, N_{\text{crit}}\}$ and the set of edges $(\mathcal{V} \setminus \mathcal{Q}) \times \mathcal{V}$ where $\mathcal{Q} = \{1, \dots, N_{\text{targ}}\}$.

Let us consider a set of particular edges $\tilde{\mathcal{E}} \subset (\mathcal{V} \setminus \mathcal{Q}) \times \mathcal{V}$ such that

$$\tilde{\mathcal{E}} := \{i \rightarrow j \in (\mathcal{V} \setminus \mathcal{Q}) \times \mathcal{V} : i \neq j, \text{cl Attr}(\mathcal{C}_i) \cap \text{cl Attr}(\mathcal{C}_j) \neq \emptyset\}. \quad (\text{B.4.30})$$

Note that by [Lemma B.1.9](#), we have that $\text{Attr}(\mathcal{C}_i)$, $i \in \{1, \dots, N_{\text{crit}}\}$ is a partition of $\mathbb{R}^d$. Therefore, $(\mathcal{V}, \tilde{\mathcal{E}})$ still satisfies that, from every $i \in \mathcal{V}$ there is a path that leads to $\mathcal{Q}$.

Lemma B.4.5. *For any $i \rightarrow j \in \tilde{\mathcal{E}}$, let us define*

$$x_{i,j} \in \arg \min_{x \in \text{cl Attr}(\mathcal{C}_i) \cap \text{cl Attr}(\mathcal{C}_j)} f(x). \quad (\text{B.4.31})$$

Then, under [Assumption B.23](#), if X is large enough so that it contains the bounded set

$$\left\{x \in \mathbb{R}^d : f(x) \leq \max_{i \in \mathcal{V} \setminus \mathcal{Q}, j \in \mathcal{V}, i \neq j} f(x_{i,j}) + 1\right\}, \quad (\text{B.4.32})$$

for any $i \rightarrow j \in \tilde{\mathcal{E}}$, it holds that

$$C_{i,j} \leq \bar{V}(x_{i,j}) - \bar{V}(\mathcal{C}_i). \quad (\text{B.4.33})$$

Proof. Note that the coercivity of f ([Assumption B.1](#)) ensure that $x_{i,j}$ is well-defined and the set of (B.4.32) is bounded.

Take $i \in \mathcal{V} \setminus \mathcal{Q}$, $j \in \mathcal{V}$, $i \neq j$. We first consider the case where $\nabla f(x_{i,j}) \neq 0$. Fix $\varepsilon > 0$.

By [Lemma B.3.22](#), we have that there exists $\delta_0 > 0$ such that, for any $\delta \in (0, \delta_0)$,

$$C_{i,j} \leq C_{i,j}^\delta + \varepsilon. \quad (\text{B.4.34})$$

At the potential cost of reducing $\delta_0 > 0$, assume that, for any $x \in \mathbb{B}(x_{i,j}, \delta_0/2)$, it holds that

$$a \leq \|\nabla f(x)\| \leq b \quad \text{and} \quad \bar{\sigma}_\infty^2(f(x)) \geq c, \quad (\text{B.4.35})$$

for some $0 < a \leq b$ and $c > 0$. Moreover, by continuity of f and $\bar{V}$, also assume that δ_0 is small enough so that, for any $x \in \mathbb{B}(x_{i,j}, \delta_0/2)$, it holds that

$$\bar{V}(x) \leq \bar{V}(x_{i,j}) + \varepsilon \quad (\text{B.4.36})$$

and,

$$f(x) \leq f(x_{i,j}) + 1. \quad (\text{B.4.37})$$

Finally, also assume that δ_0 is small enough such that, for any $x \in \mathcal{U}_\delta(\mathcal{C}_i)$, it holds that

$$\bar{V}(x) \geq \bar{V}(\mathcal{C}_i) - \varepsilon. \quad (\text{B.4.38})$$

Fix $\delta \in (0, \delta_0)$ such that $\delta \leq \varepsilon$. Now, since $x_{i,j} \in \text{cl Attr}(\mathcal{C}_i) \cap \text{cl Attr}(\mathcal{C}_j)$, there exist $x_i \in \text{Attr}(\mathcal{C}_i) \cap \mathbb{B}(x_{i,j}, \delta/2)$ and $x_j \in \text{Attr}(\mathcal{C}_j) \cap \mathbb{B}(x_{i,j}, \delta/2)$. By construction, the gradient flows started at x_i and x_j converge respectively to $\mathcal{C}_i$ and $\mathcal{C}_j$. Hence, there exists $T \geq 1$ integer such that $\Theta_{T-\alpha/2}(x_i) \in \mathcal{U}_\delta(\mathcal{C}_i)$ and $\Theta_{T-\alpha/2}(x_j) \in \mathcal{U}_\delta(\mathcal{C}_j)$ for all $\alpha := \delta/a$.

Let us now estimate the action costs of the paths $(\Theta_{T-\alpha/2-t}(x_i))_{t \in [0, T-\alpha/2]}$ and $(\Theta_t(x_j))_{t \in [0, T-\alpha/2]}$. By (B.4.37) and definition of the flow, the whole path $(\Theta_{T-\alpha/2-t}(x_i))_{t \in [0, T-\alpha/2]}$ is contained in (B.4.32) and therefore in X . Therefore, by Assumption B.23, we have that

$$\begin{aligned} \mathcal{S}_{T-\alpha/2}(\Theta_{T-\alpha/2-\cdot}(x_i)) &= \int_0^{T-\alpha/2} \mathcal{L}(\Theta_{T-\alpha/2-t}(x_i), \nabla f(\Theta_{T-\alpha/2-t}(x_i))) dt \\ &= \int_0^{T-\alpha/2} \tilde{\mathcal{L}}(\Theta_{T-\alpha/2-t}(x_i), 2\nabla f(\Theta_{T-\alpha/2-t}(x_i))) dt \\ &\leq \int_0^{T-\alpha/2} \frac{4\|\nabla f(\Theta_{T-\alpha/2-t}(x_i))\|^2}{2\bar{\sigma}^2(f(\Theta_{T-\alpha/2-t}(x_i)))} dt, \end{aligned} \quad (\text{B.4.39})$$

where we used Assumption B.23 in the last inequality. Rewriting this integral we obtain,

$$\begin{aligned} \mathcal{S}_{T-\alpha/2}(\Theta_{T-\alpha/2-\cdot}(x_i)) &\leq \int_0^{T-\alpha/2} \frac{2\langle \nabla f(\Theta_{T-\alpha/2-t}(x_i)), \dot{\Theta}_{T-\alpha/2-t}(x_i) \rangle}{\bar{\sigma}^2(f(\Theta_{T-\alpha/2-t}(x_i)))} dt \\ &= \int_0^{T-\alpha/2} 2\langle \nabla \bar{V}(\Theta_{T-\alpha/2-t}(x_i)), \dot{\Theta}_{T-\alpha/2-t}(x_i) \rangle dt \\ &= \bar{V}(x_i) - \bar{V}(\Theta_{T-\alpha/2}(x_i)) \\ &\leq \bar{V}(x_{i,j}) - \bar{V}(\mathcal{C}_i) + 2\varepsilon, \end{aligned} \quad (\text{B.4.40})$$

by (B.4.36) and (B.4.38).

For the path $(\Theta_t(x_j))_{t \in [0, T-\alpha/2]}$, since it is a trajectory of the gradient flow, we have that

$$\mathcal{S}_{T-\alpha/2}(\Theta_{\cdot}(x_j)) = 0. \quad (\text{B.4.41})$$

Finally, let us consider the path $(\gamma_t)_{t \in [0, \alpha]}$ defined by $\gamma_t = x_i + \alpha^{-1}t(x_j - x_i)$. We have $\dot{\gamma}_t = \alpha^{-1}(x_j - x_i)$ so that $\|\dot{\gamma}_t\| \leq \delta/\alpha = a \leq \|\nabla f(\gamma_t)\|$ by (B.4.35). Moreover, (B.4.37) ensures that $\gamma_t \in X$ for all $t \in [0, \alpha]$.

Therefore, by Assumption B.23, we have that

$$\begin{aligned} \mathcal{S}_\alpha(\gamma) &= \int_0^\alpha \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt \\ &= \int_0^\alpha \tilde{\mathcal{L}}(\gamma_t, \nabla f(\gamma_t)\dot{\gamma}_t) dt \\ &\leq \int_0^\alpha \frac{\|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2}{2\bar{\sigma}^2(f(\gamma_t))} dt \\ &\leq \alpha \frac{(a+b)^2}{2c} \\ &= \frac{\delta(a+b)^2}{2ac} \\ &\leq \frac{\varepsilon(a+b)^2}{2ac}. \end{aligned} \quad (\text{B.4.42})$$

Putting the three paths $(\Theta_{T-\alpha/2-t}(x_i))_{t \in [0, T-\alpha/2]}$, $(\gamma_t)_{t \in [0, \alpha]}$ and $(\Theta_t(x_j))_{t \in [0, T-\alpha/2]}$ together, we obtain a path that goes from $\mathcal{U}_\delta(\mathcal{C}_i)$ to $\mathcal{U}_\delta(\mathcal{C}_j)$ in integer time $2T$ and whose action cost is upper-bounded by

$$\bar{V}(x_{i,j}) - \bar{V}(\mathcal{C}_i) + 2\varepsilon + \frac{\varepsilon(a+b)^2}{2ac}. \quad (\text{B.4.43})$$

Combined with (B.4.34), this yields

$$C_{i,j} \leq \bar{V}(x_{i,j}) - \bar{V}(C_i) + 3\varepsilon + \frac{\varepsilon(a+b)^2}{2ac}, \quad (\text{B.4.44})$$

which yields the result.

Let us now briefly discuss the case where $\nabla f(x_{i,j}) = 0$. In this case, $x_{i,j}$ belongs to some $\mathcal{C}_l$, $l \in \mathcal{V}$. To adapt the proof above, it suffices to require that x_i and x_j are close enough to $x_{i,j}$ so that they belong to $\mathcal{W}_\varepsilon(\mathcal{C}_l)$ (give by Lemma B.1.7) and to use Lemma B.1.8 to build a path from x_i to x_j with cost at most ε . ■

We provide an iterated version of Lemma B.4.5 that will be convenient in the following.

Corollary B.4.6. *For any $i \in \mathcal{V} \setminus \mathcal{Q}$, $j \in \mathcal{V}$, define*

$$\bar{C}_{i,j} := \min \left\{ \sum_{n=0}^{N-1} \bar{V}(x_{j_n, j_{n+1}}) - \bar{V}(\mathcal{C}_{j_n}) : j_0 = i, j_N = j, j_n \rightarrow j_{n+1} \in \tilde{\mathcal{E}} \right\}. \quad (\text{B.4.45})$$

Then, under the same assumptions as Lemma B.4.5, it holds that

$$C_{i,j} \leq \bar{C}_{i,j}. \quad (\text{B.4.46})$$

Let us now specify how big enough we will need the set X from Assumption B.23 to be.

Lemma B.4.7. *Define*

$$C_\infty := \max_{i,j \in \mathcal{V} \setminus \mathcal{Q}} C_{i,j} + 1 < +\infty. \quad (\text{B.4.47})$$

Then, if X is large enough so that it contains the bounded set

$$\left\{ x \in \mathbb{R}^d : V_\infty(x) \leq \sup_{i \in \mathcal{V} \setminus \mathcal{Q}, x \in \mathcal{C}_i} \bar{V}(x) + C_\infty \right\}, \quad (\text{B.4.48})$$

then, for any $i, j \in \mathcal{V} \setminus \mathcal{Q}$, it holds that $C_{i,j}$ is equal to

$$\inf \{ \mathcal{S}_T(\gamma) : \gamma \in \mathcal{C}([0, T]), T \in \mathbb{N}, \gamma_0 \in \mathcal{C}_i, \gamma_T \in \mathcal{C}_j, \{\gamma(t) : t \in [0, T]\} \subset X \}. \quad (\text{B.4.49})$$

Proof. Note that C_∞ is finite by Assumption B.17. The rest of the result is then a direct consequence of Lemma B.3.16. ■

We can now combine Lemma B.4.4 with Lemma B.4.7 to obtain the following result.

Lemma B.4.8. *Under Assumption B.23, if X is large enough so that it satisfies the requirement of Lemma B.4.7, then for any $i, j \in \mathcal{V} \setminus \mathcal{Q}$, $i \neq j$, it holds that*

$$C_{j,i} - C_{i,j} \leq \Delta C_{j,i} + (1 - \Delta)(\bar{V}_i - \bar{V}_j), \quad (\text{B.4.50})$$

where $\Delta \in [0, 1]$ satisfies

$$\Delta \geq \sup_{x \in X} 1 - \frac{\sigma^2(f(x))}{\bar{\sigma}^2(f(x))}. \quad (\text{B.4.51})$$

Proof. Lemma B.4.7 allows us to consider only paths that are contained in X to compute $C_{i,j}$. Taking the infimum over such paths in Lemma B.4.4 yields:

$$C_{i,j} \geq (1 - \Delta)(C_{j,i} + (\bar{V}_j - \bar{V}_i)). \quad (\text{B.4.52})$$

Rearranging the terms, we obtain the result. $\blacksquare$

To simplify the notation for the following lemma, let us define, for $i \in \mathcal{V} \setminus \mathcal{Q}$, $j \in \mathcal{V}$, $i \neq j$,

$$\bar{V}_{i,j} := \bar{V}(x_{i,j}) \quad \text{and} \quad \bar{V}_j := \bar{V}(C_j). \quad (\text{B.4.53})$$

Lemma B.4.9. Assume that Assumption B.23 holds with X satisfying the assumptions of Lemma B.4.5 and Lemma B.4.7.

For any $i_0 \in \mathcal{V} \setminus \mathcal{Q}$, $g \in \mathcal{G}(i_0 \rightarrow \mathcal{Q})$, there exists $i \in \mathcal{V} \setminus \mathcal{Q}$ such that there is a path from i_0 to i in g and there is $g' \in \mathcal{G}$ such that the cost difference between g and g'

$$\sum_{k \rightarrow l \in g'} C_{k,l} - \sum_{k \rightarrow l \in g} C_{k,l} \quad (\text{B.4.54})$$

is upper-bounded by

$$\min \left\{ \max_{l=0, \dots, p-1} \max \left(\bar{V}_{j_l, j_{l+1}} - \bar{V}_{j_l}, \bar{V}_{j_l, j_{l+1}} - \bar{V}_i \right) + 2\Delta(N_{\text{crit}}^2 + 1) \sup_X \bar{V} : \right. \\ \left. j_0 = i, j_p \in \mathcal{Q}, j_l \in \mathcal{V} \setminus \mathcal{Q}, j_l \rightarrow j_{l+1} \in \tilde{\mathcal{E}}, l = 0, \dots, p-1 \right\}, \quad (\text{B.4.55})$$

where $\Delta \in [0, 1]$ satisfies

$$\Delta \geq \sup_{x \in X} 1 - \frac{\sigma^2(f(x))}{\bar{\sigma}^2(f(x))}. \quad (\text{B.4.56})$$

Proof. By definition of $\mathcal{G}(i_0 \rightarrow \mathcal{Q})$ (Appendix B.3.1), there exists $i \in \mathcal{V} \setminus \mathcal{Q}$, $j \in \mathcal{V}$, $i \neq j$ such that if we were to add the edge $i \rightarrow j$ to g , it would be a $\mathcal{Q}$ -graph. In other words, $i \rightarrow j$ is not in g and there are paths from $i_0 \rightarrow i$ and $j \rightarrow \mathcal{Q}$ in g .

Now, take consider a path that goes from i to j with only edges from $\tilde{\mathcal{E}}$: $j_0 = i$, $j_p = j$ and $j_l \in \mathcal{V} \setminus \mathcal{Q}$, $j_l \rightarrow j_{l+1} \in \tilde{\mathcal{E}}$, $l = 0, \dots, p-1$.

Define $k \geq 0$ to be the smallest integer such that there is a path from j_k to $\mathcal{Q}$ in g . Such a k exists since there is a path from j to $\mathcal{Q}$ in g , and $k \geq 1$ since there is no path from i to $\mathcal{Q}$ in g . In particular, there is no path from j_{k-1} to $\mathcal{Q}$ in g . From the definition of the graphs $\mathcal{G}(i_0 \rightarrow \mathcal{Q})$, this implies either there is a path from i to j_{k-1} in g or a path from j_{k-1} to i in g .

- If there is a path from i to j_{k-1} in g , then we consider the graph $g' \in \mathcal{G}(\mathcal{Q})$ obtained by adding the edge $j_{k-1} \rightarrow j_k$ to g . Since $j_{k-1} \rightarrow j_k$ is in $\tilde{\mathcal{E}}$, by Lemma B.4.5, the cost of adding that edge is bounded as

$$C_{j_{k-1}, j_k} \leq \bar{V}_{j_{k-1}, j_k} - \bar{V}_{j_{k-1}}, \quad (\text{B.4.57})$$

in which case (B.4.55) is immediately satisfied.

- If there is a path from j_{k-1} to i in g , then we consider the graph $g' \in \mathcal{G}(\mathcal{Q})$ obtained by reversing the path from j_{k-1} to i in g and adding the edge $j_{k-1} \rightarrow j_k$. Let us denote by $m_0 = j_{k-1}$, $m_q = i$ and $m_l \in \mathcal{V} \setminus \mathcal{Q}$, $m_l \rightarrow m_{l+1} \in g$, $l = 0, \dots, q-1$ the path from j_{k-1} to i . By Lemma B.4.8 and

Lemma B.4.5 the cost of reversing that path and adding the edge $j_{k-1} \rightarrow j_k$ is upper-bounded by

$$\begin{aligned} \sum_{l=0}^{q-1} \left(C_{m_{l+1}, m_l} - \sum_{l=0}^{p-1} C_{m_l, m_{l+1}} \right) + C_{j_{k-1}, j_k} &\leq \sum_{l=0}^{q-1} \left(\Delta C_{m_{l+1}, m_l} + (1 - \Delta)(\bar{V}_{m_l} - \bar{V}_{m_{l+1}}) \right) + \bar{V}_{j_{k-1}, j_k} - \bar{V}_{j_{k-1}} \\ &= \sum_{l=0}^{q-1} \Delta C_{m_{l+1}, m_l} + (1 - \Delta)(\bar{V}_{j_{k-1}} - \bar{V}_i) + \bar{V}_{j_{k-1}, j_k} - \bar{V}_{j_{k-1}}. \end{aligned} \quad (\text{B.4.58})$$

Now, **Corollary B.4.6** ensures that the sum $\sum_{l=0}^{q-1} \Delta C_{m_{l+1}, m_l}$ is upper-bounded by $2N_{\text{crit}}^2 \Delta \sup_X \bar{V}$ so we obtain that the RHS of (B.4.58) is upper-bounded by

$$2N_{\text{crit}}^2 \Delta \sup_X \bar{V} + \bar{V}_{j_{k-1}, j_k} - \bar{V}_i + 2\Delta \sup_X \bar{V}, \quad (\text{B.4.59})$$

which concludes the proof. $\blacksquare$

To obtain the main result, we will require the following lemma.

Lemma B.4.10. Under **Assumption B.23**, if X is large enough so that it satisfies the assumptions of **Lemma B.4.7**, then for any $i, j \in \mathcal{V} \setminus \mathcal{Q}$, $i \neq j$, it holds that For any $i \in \mathcal{V} \setminus \mathcal{Q}$, $j \in \mathcal{V}$, $i \neq j$, it holds that

$$C_{i,j} \geq \underline{C}_{i,j} := \inf \left\{ \sup_{s < t} (\underline{V}(\gamma_t) - \underline{V}(\gamma_s)) : \gamma \in \mathcal{C}([0, T]), T \in \mathbb{N}, \gamma_0 \in \mathcal{C}_i, \gamma_T \in \mathcal{C}_j \right\}. \quad (\text{B.4.60})$$

Proof. This lemma follows the same proof as **Lemma B.3.14** with the additional use of **Lemma B.4.7** to be able to leverage **Assumption B.23**. $\blacksquare$

We are now finally ready to prove the main result.

Theorem B.11. Assume that **Assumption B.23** holds with X satisfying the assumptions of **Lemma B.4.5** and **Lemma B.4.7**.

For any $i_0 \in \mathcal{V} \setminus \mathcal{Q}$, define

$$r := \min \left\{ \sum_{k \rightarrow l \in \mathcal{G}} \bar{C}_{k,l} : g \in \mathcal{G}(i_0 \rightsquigarrow \mathcal{Q}) \right\}. \quad (\text{B.4.61})$$

there exists $i \in \mathcal{V} \setminus \mathcal{Q}$ such that $\underline{C}_{i_0, i} \leq r$ such that $E(\mathcal{Q} | i_0)$ is upper-bounded by

$$\begin{aligned} \min \left\{ \max_{l=0, \dots, p-1} \max \left(\bar{V}_{j_l, j_{l+1}} - \bar{V}_{j_l}, \bar{V}_{j_l, j_{l+1}} - \bar{V}_i \right) + 2\Delta(N_{\text{crit}}^2 + 1) \sup_X \bar{V} : \right. \\ \left. j_0 = i, j_p \in \mathcal{Q}, j_l \in \mathcal{V} \setminus \mathcal{Q}, j_l \rightarrow j_{l+1} \in \tilde{\mathcal{E}}, l = 0, \dots, p-1 \right\}, \end{aligned} \quad (\text{B.4.62})$$

where $\Delta \in [0, 1]$ satisfies

$$\Delta \geq \sup_{x \in X} 1 - \frac{\sigma^2(f(x))}{\bar{\sigma}^2(f(x))}. \quad (\text{B.4.63})$$

Proof. By Lemma B.3.23, $E(\mathcal{Q} | i_0)$ is upper-bounded as

$$E(\mathcal{Q} | i_0) \leq \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(\mathcal{Q}) \right\} - \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(i_0 \not\rightarrow \mathcal{Q}) \right\}. \quad (\text{B.4.64})$$

By Corollary B.4.6, we have that

$$\min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(i_0 \not\rightarrow \mathcal{Q}) \right\} \leq \min \left\{ \sum_{k \rightarrow l \in g} \bar{C}_{k,l} : g \in \mathcal{G}(i_0 \not\rightarrow \mathcal{Q}) \right\} = r. \quad (\text{B.4.65})$$

Hence, take $g \in \arg \min \left\{ \sum_{k \rightarrow l \in g} C_{k,l} : g \in \mathcal{G}(\mathcal{Q} \rightarrow \mathcal{Q}) \right\}$ and it must satisfy

$$\sum_{k \rightarrow l \in g} C_{k,l} \leq r. \quad (\text{B.4.66})$$

Apply Lemma B.4.9 to obtain that there exists $i \in \mathcal{V} \setminus \mathcal{Q}$ such that there is a path from i_0 to i in g and there is $g' \in \mathcal{G}$ such that the cost difference between g and g' is bounded by (B.4.62). Combined with (B.4.64), this yields that $E(\mathcal{Q} | i_0)$ is upper-bounded by (B.4.62).

Finally, denote by $j_0 = i_0$, $j_p = i$ and $j_l \in \mathcal{V} \setminus \mathcal{Q}$, $j_l \rightarrow j_{l+1} \in g$, $l = 0, \dots, p-1$ the path from i_0 to i in g . (B.4.66) ensures that

$$r \geq \sum_{l=0}^{p-1} C_{j_l, j_{l+1}} \geq \underline{C}_{i_0, i}, \quad (\text{B.4.67})$$

which, combined with Lemma B.4.10, implies that $\underline{C}_{i_0, i} \leq r$. $\blacksquare$

B.5 AUXILIARY RESULTS

B.5.1 Truncated Gaussian distribution

Let us restate two technical lemmas from (Azizian et al., 2024b) on which we rely. The first one provides bounds on the Hamiltonian and Lagrangian of a truncated Gaussian distribution.

Lemma B.5.1 ((Azizian et al., 2024b, Lem. F.2)). *Consider $X \sim \mathcal{N}(0, \sigma^2 I_d)$ a multivariate Gaussian distribution with $\sigma^2 > 0$. For $R > 0$, define the truncated Gaussian r.v. X_R by conditioning X to the ball $\bar{\mathbb{B}}(0, R)$. Define*

$$\bar{\mathcal{H}}(p) := \log \mathbb{E} \left[e^{\langle p, X_R \rangle} \right] \quad (\text{B.5.1})$$

$$\bar{\mathcal{L}}(p) := \bar{\mathcal{H}}^*(p), \quad (\text{B.5.2})$$

and

$$E(\sigma^2, R) := e^{-\frac{R^2}{16\sigma^2}} 2^{d+3} (d+1) \quad (\text{B.5.3})$$

and assume that $R > 0$ satisfies

$$R \geq 4\sigma \sqrt{(d+3) \log 2 + \log(d+1)}. \quad (\text{B.5.4})$$

Then, for any $p \in \mathbb{R}^d$ such that $\|p\| \leq \frac{R}{2\sigma^2}$, $v \in \mathbb{R}^d$ such that $\|v\| \leq \frac{R}{4}$, it holds that

$$\left(1 - E(\sigma^2, R)\right) \frac{\sigma^2 \|p\|^2}{2} \leq \tilde{\mathcal{H}}(p) \leq \left(1 + E(\sigma^2, R)\right) \frac{\sigma^2 \|p\|^2}{2} \quad (\text{B.5.5})$$

$$\left(1 - 2E(\sigma^2, R)\right) \frac{\|v\|^2}{2\sigma^2} \leq \tilde{\mathcal{L}}(v) \leq \left(1 + 2E(\sigma^2, R)\right) \frac{\|v\|^2}{2\sigma^2}. \quad (\text{B.5.6})$$

We will also require the following technical lemma.

Lemma B.5.2 ((Azizian et al., 2024b, Lem. F.3)). Consider $v, w \in \mathbb{R}^d$ such that $0 < \|w\| \leq \frac{\mu R}{2}$ for some $R, \mu > 0$. Define,

$$f(u) = \sup_{p \in \mathbb{R}^d: \|p\| \leq R} \langle p, u \rangle - \frac{\mu}{2} \|p\|^2, \quad (\text{B.5.7})$$

then, with $\lambda = \frac{\|v\|}{\|w\|}$,

$$\lambda f\left(\frac{v}{\lambda} + w\right) \leq f(v + w). \quad (\text{B.5.8})$$

REGULARIZATION FOR WASSERSTEIN DISTRIBUTIONALLY ROBUST OPTIMIZATION

This appendix reproduces [Azizian et al. \(2023a\)](#) in full. Compared to the entropic regularization treated in § 6.4, this paper covers two additional aspects: (i) general doubly-regularized WDRO with arbitrary convex regularizers (beyond the entropic case), and (ii) unbounded sample spaces (beyond the compact setting of [Assumption 6.1](#)).

C.1 INTRODUCTION

OT has a long history and exciting recent developments, notably around applications in machine learning and data science; we refer to the monographs [Villani \(2003\)](#), [Santambrogio \(2015\)](#), [Peyré et al. \(2019\)](#), and [Merigot and Thibert \(2021\)](#). One of the key technical properties at the core of recent success of OT in these applications is the use of regularization, and specifically entropic regularization, opening the way to efficient computational schemes (see e.g., [Cuturi \(2013\)](#)) to get theoretically-grounded approximations of the Wasserstein distances.

DRO has recently been formulated using OT metrics and has proven to be useful in machine learning (see e.g., [Kuhn et al. \(2019\)](#)). But regularization has still to be studied and used in this context. In this paper, we propose a study of regularization in WDRO, inspired from several recent developments in OT, namely [Genevay et al. \(2016\)](#), [Carlier et al. \(2017\)](#), [Genevay et al. \(2019\)](#), and [Paty and Cuturi \(2020b\)](#).

C.1.1 *Distributionally robust optimization with Wasserstein neighborhoods*

DRO is a popular approach in optimization under uncertainty. We briefly present here the ideas and the notation that we will use; we refer to the celebrated paper [Esfahani and Kuhn \(2018\)](#) and the survey paper [Kuhn et al. \(2019\)](#) for more details and for applications in machine learning.

Standard approaches in stochastic optimization consider minimizing the expectation of a random loss with respect to some input distribution or available data points: For an objective $f_\theta : \Xi \rightarrow \mathbb{R}$ defined on a sample space Ξ and depending on parameters $\theta \in \Theta$, this consists in considering

$$\min_{\theta \in \Theta} \mathbb{E}_{\xi \sim P}[f_\theta(\xi)]. \quad (\text{C.1.1})$$

Here P is a fixed probability distribution on Ξ ; in practice, it is typically an empirical distribution $P = \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$ stemming from data samples $(\xi_i)_{i=1}^n$.

A DRO counterpart of this problem is to minimize the expectation of the loss with respect to a set of probability distributions close to P . More precisely, we choose a neighborhood $\mathcal{U}(P)$ of P (called the ambiguity set or the distributional uncertainty region) in the set of probabilities measures on Ξ , denoted by $\mathcal{P}(\Xi)$ and consider the worst possible expectation of the objective in this neighborhood. The resulting problem is thus of the form

$$\min_{\theta \in \Theta} F(\theta) \quad \text{where} \quad F(\theta) := \sup_{Q \in \mathcal{U}(P)} \mathbb{E}_{\xi \sim Q}[f_{\theta}(\xi)]. \quad (\text{C.1.2})$$

A natural way to define the ambiguity set $\mathcal{U}(P)$ is to consider a ball centered at P with radius $\rho > 0$ controlling the required level of robustness. When using the Wasserstein distance to define the ball, this gives so-called Wasserstein DRO problems (WDRO).

For a cost function $c : \Xi \times \Xi \rightarrow \mathbb{R}_+$, the Wasserstein distance between two probability distributions $P, Q \in \mathcal{P}(\Xi)$ is defined as

$$W_c(P, Q) = \inf \{ \mathbb{E}_{(\xi, \zeta) \sim \pi} c(\xi, \zeta) : \pi \in \mathcal{P}(\Xi \times \Xi), \pi_1 = P, \pi_2 = Q \}, \quad (\text{C.1.3})$$

where π_1 and π_2 denote the first and second marginals of the coupling probability, or transport plan, π defined on $\Xi \times \Xi$. A WDRO problem thus has the form (C.1.2) with the ambiguity set

$$\mathcal{U}(P) = \{Q \in \mathcal{P}(\Xi) : W(P, Q) \leq \rho\}. \quad (\text{C.1.4})$$

When the the objective f_{θ} exhibits a simple structure, this problem reformulates as a tractable convex optimization problem; see e.g., [Kuhn et al. \(2019\)](#). This is exploited in several applications, for instance: logistic regression (see e.g., [Yu et al. \(2022\)](#)), support vector machines (see e.g., [Shafieezadeh-Abadeh et al. \(2019a\)](#)), or ℓ^1 -regression (see e.g., [\(Chen and Paschalidis, 2020\)](#)). Another argument supporting a WDRO approach for machine learning applications is that it provides generalization guarantees, see e.g., [Esfahani and Kuhn \(2018\)](#), [An and Gao \(2021\)](#).

c.1.2 Contributions, related works, and outline

In this paper we study regularization in the context of Wasserstein distributionally robust optimization. First, we propose a unified framework for double regularization of the WDRO objective function (both in the objective and in the constraint). We then provide a strong duality result with general convex regularization functions. This result can be seen as the analogue for WDRO of the general result of [Paty and Cuturi \(2020b\)](#) for OT. Second, we refine our analysis in the case of the entropic regularization and obtain an explicit expression for the dual problem. Furthermore, we provide approximation guarantees when the regularization parameters are driven to 0, adapting the reasoning from [Carlier et al. \(2017\)](#). These results can be seen as analogues for WDRO of results of [Genevay et al. \(2019\)](#) for the entropic-regularized OT. Note however, that the reasoning and results used OT do not directly apply in the context of WDRO. Indeed, in OT, the entropic regularization depends on (the product of) the two marginal distributions whereas, in WDRO, the second marginal is not fixed but rather an optimization variable, which makes the analysis different and more involved. This distinction is further discussed in the beginning of [Appendix C.3](#).

Up to our knowledge, regularization in the context of WDRO has not been investigated on its own, as we do in this paper.¹ Nevertheless, let us mention the two recent papers related to our developments: [Blanchet and Kang \(2020\)](#) and [Wang et al. \(2021\)](#). In [Blanchet and Kang \(2020\)](#), an entropic smoothing of a specific WDRO dual function is introduced and used for computational purposes. Such dual smoothing implicitly corresponds to a regularization of the associated primal problem, but this link is not formally made. In contrast, the preprint [Wang et al. \(2021\)](#) (which appeared while we were preparing this manuscript) shares similar spirit as our work. The entropic regularization of WDRO is proposed and analyzed, with a special focus on computational aspects. This is complementary to our work which provides a theoretical study of general regularizations as well as approximation guarantees for the entropic regularization. We will come back more precisely on the connections between our results and those of [Blanchet and Kang \(2020\)](#) and [Wang et al. \(2021\)](#) in [Remark 9](#).

The outline of this paper is the following. The introduction ends below with the definition of the framework of this paper. In [Appendix C.2](#), we provide a duality result for a general double regularization, together with an illustration in the case when the transport cost is used a regularization function. In [Appendix C.3](#), we focus our analysis to the entropic regularization to get refined expressions and an explicit control of the quality of the approximation of the underlying WDRO problem.

c.1.3 Set-up, notation, and assumptions

The framework of this paper is the following. With Ξ a subset of $\mathbb{R}^d$, P a reference probability distribution over Ξ , and $f: \Xi \rightarrow \mathbb{R}$ the underlying objective function (we drop the dependence in θ to simplify the notation), we consider the sup problem in the objective function of [\(C.1.2\)](#) with the Wasserstein ball of radius ρ as an ambiguity set [\(C.1.4\)](#). Our objective thus writes:

$$\sup\{\mathbb{E}_{\xi \sim Q} f(\xi) : Q \in \mathcal{P}(\Xi), W(P, Q) \leq \rho\}. \quad (\text{C.1.5})$$

We reformulate the above problem, in a concise way, using only couplings as

$$\sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi): \mathbb{E}_{\pi} c \leq \rho} \mathbb{E}_{\pi_2} f \quad (\text{WDRO})$$

where $\mathcal{P}_P(\Xi \times \Xi)$ is the set of probability distributions on $\Xi \times \Xi$ having P as a first marginal

$$\mathcal{P}_P(\Xi \times \Xi) := \{\pi \in \mathcal{P}(\Xi \times \Xi) : \pi_1 = P\}. \quad (\text{C.1.6})$$

When the space Ξ is compact, we have that the topological dual of $\mathcal{C}(\Xi^2)$, the set of continuous functions on $\Xi \times \Xi$, is exactly $\mathcal{M}(\Xi^2)$, the set of finite signed measures over $\Xi \times \Xi$ by the Riesz representation ([Rudin, 1987](#), Th. 2.14). We denote by $\langle \cdot, \cdot \rangle$ the duality pairing between $\mathcal{C}(\Xi \times \Xi)$ and $\mathcal{M}(\Xi \times \Xi)$:

$$\begin{cases} \mathcal{M}(\Xi \times \Xi) \times \mathcal{C}(\Xi \times \Xi) & \longrightarrow \mathbb{R} \\ (\pi, \varphi) & \longmapsto \langle \pi, \varphi \rangle := \int \varphi \, d\pi. \end{cases} \quad (\text{C.1.7})$$

¹ Let us mention that studying “regularization for WDRO” as we propose here should not be confused with studying “the regularizing effect of WDRO on learning problems”, which is a separate field of study (see e.g., [Blanchet et al. \(2019\)](#), [Shafieezadeh-Abadeh et al. \(2019a\)](#))

When establishing general duality results, we will also make a constant use of the convex conjugate of a function $F: \mathcal{C}(\Xi \times \Xi) \rightarrow \mathbb{R} \cup \{+\infty\}$ defined as

$$F^*: \begin{cases} \mathcal{M}(\Xi \times \Xi) & \longrightarrow \mathbb{R} \cup \{+\infty\} \\ \pi & \longmapsto \sup_{\varphi \in \mathcal{C}(\Xi)} \langle \pi, \varphi \rangle - F(\varphi), \end{cases} \quad (\text{C.1.8})$$

as well as the pre-conjugate of a function $G: \mathcal{M}(\Xi \times \Xi) \rightarrow \mathbb{R} \cup \{+\infty\}$ defined as

$$G_*: \begin{cases} \mathcal{C}(\Xi \times \Xi) & \longrightarrow \mathbb{R} \cup \{+\infty\} \\ \varphi & \longmapsto \sup_{\pi \in \mathcal{M}(\Xi)} \langle \pi, \varphi \rangle - G(\pi). \end{cases} \quad (\text{C.1.9})$$

In presence of convexity, these two operations are dual one another; see e.g., [Clason and Valkonen \(2020, Rem. 5.2\)](#). More precisely, we have that $(F^*)_* = F$ when F is l.s.c., convex, and proper, and $(G_*)^* = G$ when G is weakly- $\star$ l.s.c., convex, and proper. Furthermore, the following duality result will be instrumental in our developments; it is a reformulation ([Bot et al., 2009, Th. 3.2.6](#)), adapted to our purposes.

Lemma C.1.1 (General Fenchel duality). *Consider a compact subset $\mathcal{X} \subset \mathbb{R}^d$, a function $h \in \mathcal{C}(\mathcal{X})$, and two functions $F, G: \mathcal{M}(\mathcal{X}) \rightarrow \mathbb{R} \cup \{+\infty\}$ convex weakly- $\star$ l.s.c. proper. If there exists $\varphi \in \text{dom } F_* \cap (h - \text{dom } G_*)$ such that F_* is continuous at φ , then*

$$\sup_{\pi \in \mathcal{M}(\mathcal{X})} \langle \pi, h \rangle - F(\pi) - G(\pi) = \inf_{\varphi, \psi \in \mathcal{C}(\mathcal{X}): \varphi + \psi = h} F_*(\varphi) + G_*(\psi). \quad (\text{C.1.10})$$

Proof. First, the RHS can be rewritten as

$$\inf_{\varphi, \psi \in \mathcal{C}(\mathcal{X}): \varphi + \psi = h} F_*(\varphi) + G_*(\psi) = \inf_{\varphi \in \mathcal{C}(\mathcal{X})} F_*(\varphi) + G_*(h - \varphi). \quad (\text{C.1.11})$$

Then, since F is convex weakly- $\star$ l.s.c., we get that $F = (F_*)^*$ and that F_* is proper, convex, and l.s.c.; see ([Clason and Valkonen, 2020, Rem. 5.2](#)). The same holds for G .

We can thus apply the duality result from ([Bot et al., 2009, Th. 3.2.6](#)) with F_* and $G_*(h - \cdot)$ as primal functions that are proper convex functions, and $\mathcal{C}(\mathcal{X})$ as primal space. Indeed, the regularity assumption of the lemma exactly gives the regularity condition (RC_1^{id}) of this result. Hence,

$$\inf_{\varphi \in \mathcal{C}(\mathcal{X})} F_*(\varphi) + G_*(h - \varphi) = \sup_{\pi \in \mathcal{M}(\mathcal{X})} -(F_*)^*(-\pi) - (G_*(h - \cdot))^*(\pi) \quad (\text{C.1.12})$$

$$= \sup_{\pi \in \mathcal{M}(\mathcal{X})} -F(-\pi) - G(-\pi) - \langle \pi, h \rangle. \quad (\text{C.1.13})$$

Carrying out the change of variable $\pi \leftarrow -\pi$ then concludes the proof. $\blacksquare$

C.2 REGULARIZATION OF THE WDRO OBJECTIVE FUNCTION

In this section, we study (WDRO) objectives with additional regularization functions both in the constraints $\mathbb{E}_{\pi} c \leq \rho$ and in the objective $\mathbb{E}_{\pi_2} f$. For two

arbitrary convex functions $R, S : \mathcal{M}(\Xi^2) \rightarrow \mathbb{R} \cup \{+\infty\}$, the regularized objective we consider is

$$\sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi) : \mathbb{E}_\pi c + S(\pi) \leq \rho} \mathbb{E}_{\pi_2} f - R(\pi). \quad (\text{R-WDRO})$$

We give in [Appendix C.2.1](#) the expression of the dual of this problem, under some compactness assumptions. We specialize in [Appendix C.2.2](#) the obtained expression for two specific regularizations, namely the cost regularization and the entropic one. Finally, we present in [Appendix C.2.3](#) a general framework, beyond any compactness assumption, where the dual expression still holds.

c.2.1 Duality for regularized WDRO

In this section, we give the expression of the dual of (R-WDRO). This result is the analogue for WDRO of the main result of [Paty and Cuturi \(2020b\)](#) which gives the dual of the OT problem regularized by a general convex function. We get this dual expression under the following assumptions, providing a nice duality between objects.

Assumption C.1.

- (i) $\Xi \subset \mathbb{R}^d$ is convex and compact;
- (ii) $f : \Xi \rightarrow \mathbb{R}$ and $c : \Xi \times \Xi \rightarrow \mathbb{R}_+$ are both continuous;
- (iii) For all ξ in Ξ , $c(\xi, \xi) = 0$.

This formulation of the dual problem can also be seen as a generalization of the existing one for the non-regularized case ($R = S = 0$), see e.g., [Blanchet and Murthy \(2019\)](#), [Gao and Kleywegt \(2022\)](#). Note however that the duality results in these papers rely on weaker assumptions; in particular, Ξ is not assumed to be compact and f, c are only upper- and lower-semicontinuous respectively. In the following result, these assumptions are needed to handle general regularizations. We discuss how to alleviate them for particular regularization functions in the next section.

Theorem C.1 (Strong duality for doubly-regularized WDRO). *Let [Assumption C.1](#) hold and take two convex, proper, and weakly- $\star$ l.s.c. functions $R, S : \mathcal{M}(\Xi^2) \rightarrow \mathbb{R} \cup \{+\infty\}$, such that $R + S$ is also proper. If the primal problem (R-WDRO) is strictly feasible (i.e., if there exists $\pi \in \mathcal{P}_P(\Xi \times \Xi) \cap \text{dom } R$ such that $\mathbb{E}_\pi c + S(\pi) < \rho$), then we have*

$$\text{val}(\text{R-WDRO}) = \inf_{\lambda \geq 0} \inf_{\varphi \in \mathcal{C}(\Xi^2)} \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda c(\xi, \zeta) - \varphi(\xi, \zeta) \right] + (R + \lambda S)_*(\varphi), \quad (\text{C.2.1})$$

and there exists a primal optimal solution $\pi^* \in \mathcal{P}_P(\Xi \times \Xi)$ and an optimal dual parameter $\lambda^* \geq 0$.

We prove this theorem by carefully combining two standard duality results (the Lagrangian duality theorem in Banach spaces [Peypouquet \(2015\)](#) and the Fenchel duality theorem recalled in [Lemma C.1.1](#)) with a powerful theorem for exchanging minimization/integration [Rockafellar and Wets \(1998\)](#). The latter is rarely considered in similar contexts; see [Sinha et al. \(2018, Thm. 5\)](#) for an exception.

Proof. The first step consists in applying the Lagrangian duality theorem of [Peypouquet \(2015, Thm. 3.68\)](#); let us check its assumptions. First note that

Slater's condition holds by assumption, so we only need to check that the primal problem has at least a solution. This is the case by the following arguments: (i) the problem is feasible by assumption; (ii) $\mathcal{P}(\Xi \times \Xi)$ is weakly- $\star$ sequentially compact (see e.g., [Brezis \(2010, Cor. 3.30\)](#)); (iii) the constraint set $\{\pi \in \mathcal{P}_P(\Xi \times \Xi) : \mathbb{E}_{\pi} c + S(\pi) \leq \rho\}$ is weakly- $\star$ closed (since S is weakly- $\star$ l.s.c. and the constraint $\pi_1 = P$ is weakly- $\star$ closed); and (iv) the objective $\pi \mapsto \mathbb{E}_{(\xi, \zeta) \sim \pi} [f(\zeta) - \lambda c(\xi, \zeta)] - R(\pi)$ is weakly- $\star$ u.s.c. by assumption. As a result, we have Lagrangian duality and existence of a dual solution

$$\text{val (R-WDRO)} = \inf_{\lambda \geq 0} \sup_{\pi \in \mathcal{P}(\Xi \times \Xi) : \pi_1 = P} \mathbb{E}_{(\xi, \zeta) \sim \pi} [f(\zeta) - \lambda c(\xi, \zeta)] - (R + \lambda S)(\pi) + \lambda \rho \quad (\text{C.2.2})$$

The next step is to write the inner sup as an inf. For concision, let us introduce $T_\lambda := R + \lambda S$ and $F_\lambda(\xi, \zeta) := f(\zeta) - \lambda c(\xi, \zeta)$. The sup over π , for a fixed $\lambda \geq 0$, thus writes

$$\sup_{\pi \in \mathcal{P}(\Xi \times \Xi) : \pi_1 = P} \mathbb{E}_{(\xi, \zeta) \sim \pi} [f(\zeta) - \lambda c(\xi, \zeta)] - T_\lambda(\pi) = \sup_{\pi \in \mathcal{M}(\Xi^2)} \langle \pi, F_\lambda \rangle - \iota_{\mathcal{P}_P(\Xi \times \Xi)}(\pi) - T_\lambda(\pi) \quad (\text{C.2.3})$$

where $\iota_{\mathcal{P}_P(\Xi \times \Xi)}$ is the indicator function in the sense of convex analysis, i.e., for $\pi \in \mathcal{M}(\Xi^2)$, $\iota_{\mathcal{P}_P(\Xi \times \Xi)}(\pi) = 0$ if $\pi \in \mathcal{P}(\Xi \times \Xi)$ and $\pi_1 = P$, and $+\infty$ otherwise.

Now, we want to apply the duality result of [Lemma C.1.1](#) with $F \leftarrow \iota_{\mathcal{P}_P(\Xi \times \Xi)}$, $G \leftarrow T_\lambda$, and $h \leftarrow F_\lambda$. To do so, we need to derive the (pre)conjugates of $\iota_{\mathcal{P}_P(\Xi \times \Xi)}$ and T_λ . By the disintegration theorem, any coupling $\pi(d\xi, d\zeta)$ can be written as $P(d\xi)Q(d\zeta|\xi)$ with Q a conditional probability on Ξ . Therefore,

$$\begin{aligned} (\iota_{\mathcal{P}_P(\Xi \times \Xi)})_*(\varphi) &= \sup \{ \mathbb{E}_{(\xi, \zeta) \sim \pi} \varphi(\xi, \zeta) : \pi \in \mathcal{P}(\Xi \times \Xi), \pi_1 = P \} \\ &= \sup \{ \mathbb{E}_{\xi \sim P} \mathbb{E}_{\zeta \sim Q(\cdot|\xi)} \varphi(\xi, \zeta) : Q(\cdot|\cdot) \text{ conditional probability on } \Xi \} \\ &\leq \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} \varphi(\xi, \zeta) \right]. \end{aligned} \quad (\text{C.2.4})$$

We now proceed to the reverse inequality, noting that a measurable map $\zeta : \Xi \rightarrow \Xi$ induces a conditional probability $Q(\cdot|\xi) = \delta_{\zeta(\xi)}$. Hence,

$$\begin{aligned} (\iota_{\mathcal{P}_P(\Xi \times \Xi)})_*(\varphi) &= \sup \{ \mathbb{E}_{\xi \sim P} \mathbb{E}_{\zeta \sim Q(\cdot|\xi)} \varphi(\xi, \zeta) : Q(\cdot|\cdot) \text{ conditional probability on } \Xi \} \\ &\geq \sup \{ \mathbb{E}_{\xi \sim P} \varphi(\xi, \zeta(\xi)) : \zeta : \Xi \rightarrow \Xi \text{ measurable} \}. \end{aligned} \quad (\text{C.2.5})$$

The sup is finite since Ξ is compact and φ continuous. Let us define $N : \Xi \times \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ defined as $N(\xi, \zeta) = -\varphi(\xi, \zeta)$ if $\zeta \in \Xi$ and $+\infty$ otherwise. Since φ is continuous and Ξ is closed, N is jointly l.s.c., and, as a consequence, it is a normal integrand by [Rockafellar and Wets, 1998, Ex. 14.31](#). So we can apply [Rockafellar and Wets, 1998, Thm. 14.60](#) to get that

$$\inf \{ \mathbb{E}_{\xi \sim P} -\varphi(\xi, \zeta(\xi)) : \zeta : \Xi \rightarrow \Xi \text{ measurable} \} = \mathbb{E}_{\xi \sim P} \left[\inf_{\zeta \in \Xi} -\varphi(\xi, \zeta) \right]. \quad (\text{C.2.6})$$

Inverting the signs, we have showed both upper and lower-inequalities, which means that

$$(\iota_{\mathcal{P}_P(\Xi \times \Xi)})_*(\varphi) = \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} \varphi(\xi, \zeta) \right] \quad \text{for any } \varphi \in \mathcal{C}(\Xi^2), \quad (\text{C.2.7})$$

and thus $\text{dom}(\iota_{\mathcal{P}_P(\Xi \times \Xi)})_* = \mathcal{C}(\Xi^2)$. Also, $\varphi \in \mathcal{C}(\Xi^2) \mapsto \sup_{\zeta \in \Xi} \varphi(\cdot, \zeta) \in \mathcal{C}(\Xi)$ is 1-Lipschitz w.r.t. the norm of the uniform convergence so $(\iota_{\mathcal{P}_P(\Xi \times \Xi)})_*$ is continuous on its domain.

Finally, since T_λ is convex, proper and weakly- $\star$ l.s.c., $T_{\lambda*}$ is proper and therefore $\text{dom } T_* \neq \emptyset$. Thus, [Lemma C.1.1](#) can be used and gives that

$$\sup_{\pi \in \mathcal{P}(\Xi \times \Xi): \pi_1 = P} \mathbb{E}_{(\xi, \zeta) \sim \pi} [f(\zeta) - \lambda c(\xi, \zeta)] - T(\pi) = \inf_{\varphi \in \mathcal{C}(\Xi^2)} \mathbb{E}_{\xi \sim P} \sup_{\zeta \in \Xi} f(\zeta) - \lambda c(\xi, \zeta) - \varphi(\xi, \zeta) + T_{\lambda*}(\varphi), \quad (\text{C.2.8})$$

which, combined with [\(C.2.2\)](#), leads to the claimed result. $\blacksquare$

c.2.2 Examples of regularized WDRO

As an illustration of the duality result, let us consider two cases: first, when the regularization is the transport cost itself and, second, when it is a ϕ -divergence.

When the transport cost itself is used a regularization, the expression of the dual simplifies as follow. This expression will be used in the analysis of the next section.

Corollary C.2.2 (Duality for cost-regularized WDRO). *Let [Assumption C.1](#) hold and take $\varepsilon, \delta > 0$. Then we have*

$$\sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi): \mathbb{E}_{\pi} c + \delta \mathbb{E}_{\pi} c \leq \rho} \mathbb{E}_{\pi_2} f - \varepsilon \mathbb{E}_{\pi} c = \inf_{\lambda \geq 0} \lambda \rho + \mathbb{E}_{\xi \sim P} \sup_{\zeta \in \Xi} \{f(\zeta) - (\varepsilon + (1 + \delta)\lambda)c(\xi, \zeta)\}. \quad (\text{C.2.9})$$

Proof. Define $R : \pi \in \mathcal{M}(\Xi^2) \mapsto \varepsilon \langle \pi, c \rangle$, $S : \pi \in \mathcal{M}(\Xi^2) \mapsto \delta \langle \pi, c \rangle$ which are convex, proper and weakly- $\star$ continuous by construction. The pre-conjugate of their sum is $(R + \lambda S)_* = \iota_{\{(\varepsilon + \lambda\delta)c\}}$. Moreover, the primal is strictly feasible thanks to the transport plan $\pi(d\xi, d\zeta) = P(d\xi) \delta_\xi(d\zeta)$. Thus, we can apply [Theorem C.1](#) to get the expression. $\blacksquare$

When ϕ -divergences are used as regularizations, the expression of the dual problem of [\(C.2.1\)](#) simplifies as follows. We will come back in more details in [Appendix C.3](#) on the KL-divergence, which is a popular ϕ -divergence.

Corollary C.2.3 (Duality for ϕ -divergence-regularized WDRO). *Let [Assumption C.1](#) hold and take $\varepsilon, \delta \geq 0$. Consider $\pi_0 \in \mathcal{P}(\Xi \times \Xi)$, a convex l.s.c. function $\phi : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ such that $\phi(1) = 0$ and $\phi'(\pm\infty) = \pm\infty$ and define the associated divergence for all $\pi \in \mathcal{P}(\Xi \times \Xi)$ as*

$$D_\phi(\pi|\pi_0) := \begin{cases} \int_{\Xi^2} \phi\left(\frac{d\pi}{d\pi_0}\right) d\pi_0 & \text{if } \pi \text{ is absolutely continuous w.r.t. } \pi_0 \\ +\infty & \text{otherwise.} \end{cases} \quad (\text{C.2.10})$$

Then, with $R \leftarrow \varepsilon D_\phi(\cdot|\pi_0)$ and $S \leftarrow \delta D_\phi(\cdot|\pi_0)$, if the primal ([R-WDRO](#)) is strictly feasible, its value is equal to

$$\inf_{\lambda \geq 0} \inf_{\psi \in \mathcal{C}(\Xi^2)} \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda c(\xi, \zeta) - \psi(\xi, \zeta) \right] + (\varepsilon + \lambda\delta) \int_{\Xi^2} \phi^*\left(\frac{\psi(\xi, \zeta)}{\varepsilon + \lambda\delta}\right) d\pi_0(\xi, \zeta). \quad (\text{C.2.11})$$

Proof. R and S are convex and proper by construction. Furthermore, D_ϕ admits the following variational formula, see e.g., [Agrawal and Horel \(2021, Prop. 4.2.8\)](#) applied with the pair $(L^\infty(\pi_0), \mathcal{M}(\Xi^2))$,

$$\forall \pi \in \mathcal{M}(\Xi^2), D_\phi(\pi|\pi_0) = \sup \left\{ \langle \pi, \psi \rangle - \int_{\Xi^2} \phi^* \circ \psi \, d\pi_0 : \psi \in L^\infty(\pi_0) \right\}. \quad (\text{C.2.12})$$

By Lusin's theorem, see e.g., [Rudin \(1987, Thm. 2.24\)](#), the maximization over $L^\infty(\pi_0)$ can be replaced by maximization over $\mathcal{C}(\Xi^2)$, which guarantees that D_ϕ is indeed weak- $\star$ l.s.c.. Then, we can apply [Theorem C.1](#), using that the preconjugate of D_ϕ is exactly $\psi \mapsto \int_{\Xi^2} \phi^* \circ \psi \, d\pi_0$ by [Agrawal and Horel \(2021, Prop. 4.2.6\)](#). $\blacksquare$

c.2.3 Duality under weaker assumptions

In this section, we present a generalization of [Theorem C.1](#) with weaker assumptions. The main change is that the compactness of Ξ is no longer required at the expense of a growth condition on f . Specifically, [Assumption C.1](#) is replaced by the following set of assumptions.

Assumption C.2.

- (i) $\Xi \subset \mathbb{R}^d$ is a closed set;
- (ii) $f: \Xi \rightarrow \mathbb{R}$ is measurable w.r.t. the Borel σ -algebra on Ξ ;
- (iii) $c: \Xi \times \Xi \rightarrow \mathbb{R}_+$ is continuous and, for all ξ in Ξ , $c(\xi, \xi) = 0$;
- (iv) There is some $\xi_0 \in \Xi$ such that f satisfies the growth condition²:

$$\exists \alpha > 0, \text{ s.t. } \forall \xi, \zeta \in \Xi, \quad |f(\zeta)| \leq \alpha(1 + c(\xi, \xi_0) + c(\xi, \zeta)), \quad (\text{C.2.13})$$

and the integrability condition $\mathbb{E}_{\xi \sim P}[c(\xi, \xi_0)] < +\infty$ holds.

For the sake of readability, we only state here an informal version of our result, where we omit some technical assumptions on the regularizers and the construction of the spaces involved. The formal statement ([Theorem C.6](#)), along with its proof, is provided in [Appendix C.5](#).

Theorem C.4 (Strong duality for general doubly-regularized WDRO, informal). *Let [Assumption C.2](#) hold and take two convex, proper extended-valued functions R, S which satisfy some regularity conditions. If the primal problem (R-WDRO) is strictly feasible, then we have*

$$\text{val (R-WDRO)} = \inf_{\lambda \geq 0} \inf_{\varphi \in \mathcal{X}} \lambda \rho + \mathbb{E}_{\xi \sim P} [\text{ess sup}_{\zeta \in \Xi} f(\zeta) - \lambda c(\xi, \zeta) - \varphi(\xi, \zeta)] + (R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*})_*(\varphi), \quad (\text{C.2.14})$$

where $\mathcal{X}$ is a Banach function space built from c, R and S , and the essential supremum is taken w.r.t. the Lebesgue measure on Ξ .

The gist of the proof consists in carefully crafting $\mathcal{X}$ from c, R and S and then taking advantage of the duality structure of the pair $(\mathcal{X}, \mathcal{X}^*)$. Up to the

² In the case when c is a power of a distance $c = d^p$, the growth condition is equivalent to $|f(\xi)| \leq \alpha(1 + d(\xi, \xi_0)^p)$, which is standard in unregularized WDRO, see e.g., [Gao and Kleywegt \(2022\)](#), [Blanchet et al. \(2021b\)](#)

function spaces involved, the dual expression of (C.2.14) corresponds to the one of (C.2.1). Also, the regularity conditions on R and S encompass the examples of Appendix C.2.2, i.e., the cost regularization and ϕ -divergences. Note also that the continuity assumption on the cost function c can be removed in the case of ϕ -divergences; see Appendix C.5.

C.3 ENTROPIC REGULARIZATION

In this section, we specialize and refine the study of the previous section in the case entropic regularization, i.e., when the KL divergence is used a regularizing function. This regularization is defined for two signed measures with finite variations μ, ν as

$$\text{KL}(\mu|\nu) = \begin{cases} \int \log \frac{d\mu}{d\nu} d\mu & \text{if } \mu \text{ and } \nu \text{ are non-negative and } \mu \ll \nu \\ +\infty & \text{otherwise} \end{cases}. \quad (\text{C.3.1})$$

This kind of regularization is very popular in computational OT, where it enables the derivation useful approximations of the Wasserstein distance, as, for example the so-called Sinkhorn distance:

$$\inf_{\pi \in \mathcal{P}(\Xi \times \Xi): \pi_1 = P, \pi_2 = Q} \mathbb{E}_{\pi} c + \varepsilon \text{KL}(\pi | P \otimes Q) \quad \text{for } P, Q \in \mathcal{P}(\Xi) \text{ given.} \quad (\text{C.3.2})$$

When Ξ is compact, the dual of this problem is given in Genevay et al. (2016, Prop. 2.1)

$$\sup_{f \in C(\Xi)} \mathbb{E}_P f - \varepsilon \mathbb{E}_{\xi \sim P} \log \left(\mathbb{E}_{\zeta \sim Q} e^{\frac{f(\xi) - c(\xi, \zeta)}{\varepsilon}} \right). \quad (\text{C.3.3})$$

In Appendix C.3.1, we establish a similar result for WDRO. But let us point out here a technical difficulty arising from the WDRO framework compared to OT. The KL divergence (C.3.2) is taken w.r.t. the measure $P \otimes Q$, which does not restrict the set of feasible transport plans (if $\pi \in \mathcal{P}(\Xi \times \Xi)$ satisfies $\pi_1 = P$ and $\pi_2 = Q$, π is indeed absolutely continuous w.r.t. $P \otimes Q$). In WDRO however, only one marginal of the transport plans π is fixed and the support of the optimal coupling can be arbitrary, so that the same regularization as in (C.3.2) cannot directly be used.

We thus propose to regularize (WDRO) with KL using a base coupling π_0 with first marginal $(\pi_0)_1 = P$ and consider

$$\sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi): \mathbb{E}_{\pi} c + \delta \text{KL}(\pi | \pi_0) \leq \rho} \mathbb{E}_{\pi_2} f - \varepsilon \text{KL}(\pi | \pi_0). \quad (\text{E-WDRO})$$

The choice of π_0 restricts the set of transport plans to those that are absolutely continuous w.r.t. π_0 . We see in Appendix C.3.2 that this restriction has a very limited impact since a natural choice of π_0 still provides a good approximations of the original problem (WDRO) when the regularization parameters ε, δ vanish.

C.3.1 Duality for the entropy-regularized problem

We derive here a duality theorem for (E-WDRO) involving the KL regularization with an arbitrary base coupling $\pi_0 \in \mathcal{P}_P(\Xi \times \Xi)$. The result naturally involves the same features as in (C.3.3).

Theorem C.1 (Strong duality for entropy-regularized problems). *Let [Assumption C.1](#) hold, take $\varepsilon, \delta > 0$, and fix an arbitrary $\pi_0 \in \mathcal{P}_P(\Xi \times \Xi)$. If the primal problem (E-WDRO) is strictly feasible (i.e., if there exists π such that $\mathbb{E}_{\pi} c + \delta \text{KL}(\pi | \pi_0) < \rho$), then*

$$\text{val}(\text{E-WDRO}) = \inf_{\lambda \geq 0} \lambda \rho + (\varepsilon + \lambda \delta) \mathbb{E}_{\xi \sim P} \log \left(\mathbb{E}_{\zeta \sim \pi_0(\cdot | \xi)} e^{\frac{f(\zeta) - \lambda c(\xi, \zeta)}{\varepsilon + \lambda \delta}} \right), \quad (\text{C.3.4})$$

and there exists a primal optimal solution $\pi^* \in \mathcal{P}_P(\Xi \times \Xi)$ and an optimal dual parameter $\lambda^* \geq 0$.

The interest of using the entropy as a regularization appears when comparing (C.3.4) to the general dual (C.2.1). We see that there is no inf on φ and moreover the inner sup is replaced by a smoothed approximation (of the log-sum-exp type).

Remark 9 (Two related results and weaker assumptions). The expression (C.3.4) already appeared, for special cases, in recent articles/preprints. First, the dual problem in (C.3.4) is proposed in [Blanchet and Kang \(2020\)](#) as a smoothing technique (independently from duality and KL-regularization) in a specific context of semi-supervised learning over a finite set Ξ . Second, a very similar duality result (in the case of the KL-regularization in constraints only) appears in [\(Wang et al., 2021, Thm. 1\)](#). We note that this duality result holds under weaker assumptions: their proof is based on arguments specific to the KL divergence, whereas the one of [Theorem C.1](#) is based on [Corollary C.2.3](#) for ϕ -divergences and thus on [Theorem C.1](#) for generic regularizations.

It is possible to weaken the assumptions of [Theorem C.1](#) by using the general [Theorem C.4](#) instead of [Theorem C.1](#), which essentially replaces [Assumption C.1](#) by [Assumption C.2](#). We could also weaken the assumptions by using an alternative proof from duality formulas in variational inference; see [\(Lee, 2021, Thm. 2.1\)](#) which is itself inspired by [Boucheron et al. \(2005, Lem. 1\)](#). We do not provide the details of these two refinements here. Indeed, obtaining duality under weaker assumptions is not what matters most here since the approximation result of the next section requires a combination of compactness and continuity. ■

Proof. We start by applying [Corollary C.2.3](#) with $\phi : x \in \mathbb{R}_+ \mapsto x \log x - x + 1$ (whose conjugate is $\phi^*(y) = e^y - 1$), to get

$$\text{val}(\text{E-WDRO}) = \inf_{\lambda \geq 0} \lambda \rho + \underbrace{\inf_{\varphi \in \mathcal{C}(\Xi^2)} \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda c(\xi, \zeta) - \varphi(\xi, \zeta) \right]}_{(a)} + (\varepsilon + \lambda \delta) \langle \pi_0, e^{\frac{\varphi}{\varepsilon + \lambda \delta}} - 1 \rangle. \quad (\text{C.3.5})$$

For a fixed $\lambda \geq 0$, we can simplify the expression of the term (a) above by introducing $\tau := \varepsilon + \lambda \delta$, defining $F_\lambda : (\xi, \zeta) \mapsto f(\zeta) - \lambda c(\xi, \zeta) \in \mathcal{C}(\Xi^2)$, and carrying out the change of variable $\varphi \leftarrow F_\lambda - \varphi$. We thus obtain

$$(a) = \inf_{\varphi \in \mathcal{C}(\Xi^2)} \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda c(\xi, \zeta) - \varphi(\xi, \zeta) \right] + \tau \mathbb{E}_{\pi_0} \left[e^{\frac{\varphi}{\tau}} - 1 \right] \quad (\text{C.3.6})$$

$$= \inf_{\varphi \in \mathcal{C}(\Xi^2)} \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} \varphi(\xi, \zeta) \right] + \tau \mathbb{E}_{\pi_0} \left[e^{\frac{F_\lambda - \varphi}{\tau}} - 1 \right]. \quad (\text{C.3.7})$$

The rest of the proof is devoted to the reformulation of the term (a) expressed as (C.3.7). In order to get rid of the supremum in this expression, we restrict the minimization to $\mathcal{C}(\Xi)$ instead of $\mathcal{C}(\Xi^2)$, i.e., we consider functions $\varphi \in \mathcal{C}(\Xi^2)$ of the form $\varphi(\xi, \zeta) = g(\xi)$ with $g \in \mathcal{C}(\Xi)$:

$$\text{val (C.3.7)} \leq \inf_{g \in \mathcal{C}(\Xi)} \mathbb{E}_{\xi \sim P} [g(\xi)] + \tau \mathbb{E}_{(\xi, \zeta) \sim \pi_0} \left[e^{\frac{F_\lambda(\xi, \zeta) - g(\xi)}{\tau}} - 1 \right]. \quad (\text{C.3.8})$$

Let us prove that the inequality above is actually an equality. Fix a bivariate function $\varphi \in \mathcal{C}(\Xi^2)$ and consider the univariate function $g^\varphi := \sup_{\zeta \in \Xi} \varphi(\cdot, \zeta)$. Then, since $-\varphi(\xi, \zeta) \geq -g^\varphi(\xi)$, we have

$$\mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} \varphi(\xi, \zeta) \right] + \tau \mathbb{E}_{\pi_0} \left[e^{\frac{F_\lambda - \varphi}{\tau}} - 1 \right] \geq \mathbb{E}_{\xi \sim P} [g^\varphi(\xi)] + \tau \mathbb{E}_{(\xi, \zeta) \sim \pi_0} \left[e^{\frac{F_\lambda(\xi, \zeta) - g^\varphi(\xi)}{\tau}} - 1 \right]. \quad (\text{C.3.9})$$

However, g^φ is not continuous in general but only l.s.c., hence we cannot lower bound the RHS of (C.3.9) by the RHS of (C.3.8). To remedy this issue, we approximate g^φ with continuous functions defined for $k \geq 1$ by $g_k^\varphi(\xi) := \inf_{\zeta \in \Xi} g^\varphi(\zeta) + k \|\xi - \zeta\|$. Since g^φ is l.s.c., the functions (g_k^φ) converge pointwise to g^φ when k goes to $+\infty$, see e.g., (Rockafellar and Wets, 1998, Ex. 9.11). Moreover, these functions are uniformly bounded by $\sup_{\xi \in \Xi} |g^\varphi(\xi)|$, thus Lebesgue's dominated convergence implies that

$$\mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} \varphi(\xi, \zeta) \right] + \tau \mathbb{E}_{\pi_0} \left[e^{\frac{F_\lambda - \varphi}{\tau}} - 1 \right] \geq \mathbb{E}_{\xi \sim P} [g^\varphi(\xi)] + \tau \mathbb{E}_{(\xi, \zeta) \sim \pi_0} \left[e^{\frac{F_\lambda(\xi, \zeta) - g^\varphi(\xi)}{\tau}} - 1 \right] \quad (\text{C.3.10})$$

$$= \lim_{k \rightarrow +\infty} \mathbb{E}_{\xi \sim P} [g_k^\varphi(\xi)] + \tau \mathbb{E}_{(\xi, \zeta) \sim \pi_0} \left[e^{\frac{F_\lambda(\xi, \zeta) - g_k^\varphi(\xi)}{\tau}} - 1 \right] \quad (\text{C.3.11})$$

$$\geq \inf_{g \in \mathcal{C}(\Xi)} \mathbb{E}_{\xi \sim P} [g(\xi)] + \tau \mathbb{E}_{(\xi, \zeta) \sim \pi_0} \left[e^{\frac{F_\lambda(\xi, \zeta) - g(\xi)}{\tau}} - 1 \right]. \quad (\text{C.3.12})$$

Combining (C.3.8) and (C.3.12) gives that

$$(a) = \inf_{g \in \mathcal{C}(\Xi)} \mathbb{E}_{\xi \sim P} [g(\xi)] + \tau \mathbb{E}_{(\xi, \zeta) \sim \pi_0} \left[e^{\frac{F_\lambda(\xi, \zeta) - g(\xi)}{\tau}} - 1 \right] \quad (\text{C.3.13})$$

The final step of the proof consists in solving the above minimum over g . Indeed, the objective

$$g \mapsto \mathbb{E}_{\xi \sim P} [g(\xi)] + \tau \mathbb{E}_{(\xi, \zeta) \sim \pi_0} \left[e^{\frac{F_\lambda(\xi, \zeta) - g(\xi)}{\tau}} - 1 \right] \quad (\text{C.3.14})$$

is convex and differentiable on $\mathcal{C}(\Xi)$. As a consequence, the critical points are minimizers. The gradient at a continuous function $g \in \mathcal{C}(\Xi)$ lives in $\mathcal{M}(\Xi)$ and is given by

$$P(d\xi) - \left(\int_{\Xi} e^{\frac{F_\lambda(\xi, \zeta) - g(\xi)}{\tau}} \pi_0(d\zeta | \xi) \right) P(d\xi). \quad (\text{C.3.15})$$

Thus the continuous function

$$g^\star : \xi \mapsto \tau \log \left(\int_{\Xi} e^{\frac{F_A(\xi, \zeta)}{\tau}} \pi_0(d\zeta | \xi) \right) \quad (\text{C.3.16})$$

is a solution of (C.3.13). We get

$$\begin{aligned} (a) &= \mathbb{E}_{\xi \sim P} \left[\tau \log \left(\int_{\Xi} e^{\frac{F_A(\xi, \zeta)}{\tau}} \pi_0(d\zeta | \xi) \right) \right] + \underbrace{\tau \mathbb{E}_{\xi \sim P} \left[\int_{\Xi} e^{\frac{F_A(\xi, \zeta) - g^\star(\xi)}{\tau}} \pi_0(d\zeta | \xi) - 1 \right]}_{=0 \text{ (by nullity of the gradient in (C.3.15) for } g^\star)} \\ & \quad (\text{C.3.17}) \end{aligned}$$

which in turn gives the desired expression (C.3.4). $\blacksquare$

Remark 10 (Optimal transport plan). In the setting of Theorem C.1, the primal optimal solution $\pi^\star \in \mathcal{P}_P(\Xi \times \Xi)$ can be built explicitly from $\lambda^\star$ by taking for any $\xi \in \Xi$

$$\pi^\star(d\zeta | \xi) \propto e^{\frac{f(\zeta) - \lambda^\star c(\xi, \zeta)}{\varepsilon + \lambda^\star \delta}} \pi_0(d\zeta | \xi). \quad (\text{C.3.18})$$

Indeed, since the dual function

$$\lambda \mapsto \lambda \rho + (\varepsilon + \lambda \delta) \mathbb{E}_{\xi \sim P} \log \left(\mathbb{E}_{\zeta \sim \pi_0(\cdot | \xi)} e^{\frac{f(\zeta) - \lambda c(\xi, \zeta)}{\varepsilon + \lambda \delta}} \right) \quad (\text{C.3.19})$$

is differentiable at $\lambda^\star$, the optimality condition for (C.3.4) implies that $\mathbb{E}_{\pi^\star}[c] + \delta \text{KL}(\pi^\star | \pi_0) \leq \rho$ (with equality if and only if $\lambda^\star > 0$), i.e., that $\pi^\star$ is feasible in (E-WDRO). Then, one can check that

$$\mathbb{E}_{\pi_2^\star}[f] - \varepsilon \text{KL}(\pi^\star | \pi_0) = \lambda^\star \rho + (\varepsilon + \lambda^\star \delta) \mathbb{E}_{\xi \sim P} \log \left(\mathbb{E}_{\zeta \sim \pi_0(\cdot | \xi)} e^{\frac{f(\zeta) - \lambda^\star c(\xi, \zeta)}{\varepsilon + \lambda^\star \delta}} \right) \quad (\text{C.3.20})$$

so that, by strong duality (Theorem C.1), $\pi^\star$ is optimal for (E-WDRO).

c.3.2 Approximation error of entropy-regularized problems

In this section, we study the behavior of the approximation error of (E-WDRO) as the regularization parameters vanish to 0. To quantify this approximation, we specify the cost c and the reference measure π_0 . Specifically, we consider that the cost c is a norm to some power $p \geq 1$

$$c(\xi, \zeta) = \|\xi - \zeta\|^p \quad (\text{C.3.21})$$

and that the reference measure $\pi_0 \in \mathcal{P}_P(\Xi \times \Xi)$ is taken, for some $\sigma > 0$, as

$$\pi_0(d\xi, d\zeta) \propto P(d\xi) \mathbb{1}_{\zeta \in \Xi} e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} d\zeta. \quad (\text{C.3.22})$$

For example, when $c(\xi, \zeta) = \|\xi - \zeta\|_2^p$ with $p \in \{1, 2\}$, $\pi_0(\cdot | \xi)$ is a Laplace or a Gaussian distribution (which is easy to sample from for any $\xi \in \Xi$). We also slightly strengthen Assumption C.1 by assuming that Ξ is a convex body and that the functions are Lipschitz continuous.

Theorem C.2 (Approximation for entropic regularization). *Let the following conditions hold:*

- (i) the objective $f : \Xi \rightarrow \mathbb{R}$ and the cost $c : \Xi \times \Xi \rightarrow \mathbb{R}_+$ are Lipschitz continuous;

(ii) the cost c and the coupling π_0 are taken as (C.3.21) and (C.3.22) with $\sigma > 0$ such that $\mathbb{E}_{\pi_0} c < \rho$;

(iii) the set $\Xi \subset \mathbb{R}^d$ is compact, convex, with nonempty interior.

Then, as the regularization parameters $\varepsilon, \delta > 0$ go to zero, we have

$$0 \leq \text{val}(\text{WDRO}) - \text{val}(\text{E-WDRO}) \leq \mathcal{O}\left(d(\varepsilon + \bar{\lambda}\delta) \log \frac{1}{\varepsilon + \bar{\lambda}\delta}\right) \quad (\text{C.3.23})$$

where $\bar{\lambda} := \frac{2 \sup_{\Xi} |f|}{\rho - \mathbb{E}_{\pi_0} c}$ is an explicit dual bound.

This result is the analogue for (WDRO) of quantitative bounds for (C.3.2) established in Genevay et al. (2019) in the context of OT. The core of the proof consists in introducing a block approximation of the optimal transport plan, following Carlier et al. (2017). In our situation, the second marginal of the transport plan π is not fixed, and it is the variable of the optimization problem. Moreover we have an additional variable λ to take into account. So we have to be careful to modify the approximation scheme of Carlier et al. (2017), Genevay et al. (2019) and we introduce an auxiliary regularized problem, that can be better handled than the entropy-regularized problem. Thus the proof of Theorem C.2 requires several original steps, as described in the next section.

C.3.3 Proof of the approximation theorem

This section is devoted to the proof of Theorem C.2. In fact, we state and prove a slightly more detailed result, formalized in the next theorem, and we show afterwards how Theorem C.2 can be derived as a consequence. The following theorem can indeed be seen as a global version of Theorem C.2, with explicit constants and slightly more general assumptions. To simplify the reading, we denote by the optimal solution of the *entropy*-regularized problem.

$$F_{\rho}^{\varepsilon, \delta}(f, P) = \sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi): \mathbb{E}_{\pi} c + \delta \text{KL}(\pi | \pi_0) \leq \rho} \mathbb{E}_{\pi} f - \varepsilon \text{KL}(\pi | \pi_0). \quad (\text{C.3.24})$$

Theorem C.3 (Extended approximation theorem). *Take a radius $\rho > 0$, regularization parameters $\varepsilon, \delta > 0$, and suppose that the following conditions hold:*

- (i) *The objective $f: \Xi \rightarrow \mathbb{R}$ and the cost $c: \Xi \times \Xi \rightarrow \mathbb{R}_+$ are Lipschitz continuous, and that their respective Lipschitz constants satisfy $\varepsilon \leq \text{Lip}(f)$ and $\delta \leq \text{Lip}(c)$;*
- (ii) *The cost c and the coupling π_0 are taken as (C.3.21) and (C.3.22) with $\sigma > 0$ such that $\mathbb{E}_{\pi_0} c < \rho$;*
- (iii) *The set $\Xi \subset \mathbb{R}^d$ is compact, convex, and satisfies (for $\bar{\mathbf{B}}(\xi, \Delta)$ the ball for $\|\cdot\|$)*

$$V := \inf_{\xi \in \Xi, 0 < \Delta \leq d} \frac{\text{vol}(\Xi \cap \bar{\mathbf{B}}(\xi, \Delta))}{\Delta^d} > 0. \quad (\text{C.3.25})$$

Then, we have,

$$F_{\frac{\rho}{1+\delta/\sigma}}^{0,0}(f) - (\varepsilon + \bar{\lambda}\delta) \left(d + d \log \left(\frac{L}{(\varepsilon + \bar{\lambda}\delta)d} \right) + C + \frac{1}{\sigma} \left(\frac{(\varepsilon + \bar{\lambda}\delta)d}{L} \right)^p \right) - \frac{\varepsilon \rho}{\sigma + \delta} \quad (\text{C.3.26})$$

$$\leq F_{\rho}^{\varepsilon, \delta}(f) \leq F_{\rho}^{0,0}(f). \quad (\text{C.3.27})$$

where $\bar{\lambda} = \frac{2 \sup_{\Xi} |f|}{\rho - \mathbb{E}_{\pi_0} c}$, $L = \text{Lip}(f) + \bar{\lambda} \text{Lip}(c)$, and $C = \min \left\{ \log \frac{\text{vol}(\Xi)}{V}, \log \frac{I_\sigma}{V} \right\}$ with $I_\sigma = \sigma^{\frac{d}{p}} \int_{\mathbb{R}^d} e^{-\frac{\|\zeta\|^p}{2^{p-1}}} d\zeta$.

The proof of this result requires a few preliminary steps. First, we provide in [Lemma C.3.4](#) a simple approximation result for the *cost*-regularized problem

$$G_\rho^{\varepsilon, \delta}(f) = \sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi): \mathbb{E}_\pi c + \delta \mathbb{E}_\pi c \leq \rho} \mathbb{E}_{\pi_2} f - \varepsilon \mathbb{E}_\pi c. \quad (\text{C.3.28})$$

Next, we bound in [Lemma C.3.5](#) the dual optimal solution of the *entropy*-regularized problem $F_\rho^{\varepsilon, \delta}(f, P)$. Finally, for a fixed dual variable, we compare in [Lemma C.3.6](#) the values of the Lagrangians of the *entropy*-regularized problem and the *cost*-regularized one, which is the most technical part of the proof. After these three lemmas, we prove the approximation result in the extended version ([Theorem C.3](#)), and show how the initial version ([Theorem C.2](#)) can be derived from it.

Lemma C.3.4 (Approximation for cost-regularization). *Let [Assumption C.1](#) hold and take $\varepsilon, \delta > 0$. Then, the following bound hold*

$$G_\rho^{0,0}(f) - \frac{\varepsilon \rho}{1 + \delta} \leq G_\rho^{\varepsilon, \delta}(f) \leq G_\rho^{0,0}(f). \quad (\text{C.3.29})$$

Proof. Since the cost function is non-negative, we directly have $G_\rho^{\varepsilon, \delta}(f) \leq G_\rho^{0,0}(f)$. From [Corollary C.2.2](#), we write the cost-regularized function $G_\rho^{\varepsilon, \delta}(f)$ of [\(C.3.28\)](#) as follows

$$G_\rho^{\varepsilon, \delta}(f) = \inf_{\lambda \geq 0} \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - (\varepsilon + (1 + \delta)\lambda)c(\xi, \zeta) \right] \quad (\text{C.3.30})$$

$$= \inf_{\lambda' \geq \varepsilon} \frac{\lambda' - \varepsilon}{1 + \delta} \rho + \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda' c(\xi, \zeta) \right] \quad (\text{with } \lambda' = \varepsilon + (1 + \delta)\lambda) \quad (\text{C.3.31})$$

$$\geq \inf_{\lambda' \geq 0} \frac{\lambda' - \varepsilon}{1 + \delta} \rho + \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda' c(\xi, \zeta) \right] \quad (\text{C.3.32})$$

$$\geq \frac{-\varepsilon \rho}{1 + \delta} + \inf_{\lambda' \geq 0} \lambda' \frac{\rho}{1 + \delta} + \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda' c(\xi, \zeta) \right] \quad (\text{C.3.33})$$

$$= \sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi): \mathbb{E}_\pi c \leq \frac{\rho}{1 + \delta}} \mathbb{E}_{\pi_2} f - \frac{\varepsilon \rho}{1 + \delta} = G_\rho^{0,0}(f) - \frac{\varepsilon \rho}{1 + \delta}, \quad (\text{C.3.34})$$

where the last line follows again from [Corollary C.2.2](#) with $\varepsilon = \delta = 0$. $\blacksquare$

Lemma C.3.5 (Upper-bound on dual solutions). *Under the assumptions of [Theorem C.1](#), the optimal solution $\lambda^\star$ of the dual problem of $F_\rho^{\varepsilon, \delta}(f, P)$ is bounded as follows*

$$\lambda^\star \leq \bar{\lambda} = \frac{2 \sup_{\Xi} |f|}{\rho - \mathbb{E}_{\pi_0} c}. \quad (\text{C.3.35})$$

Proof. [Theorem C.1](#) gives the existence of $\lambda^\star$, which, by definition, minimizes

$$g: \lambda \mapsto \lambda \rho + (\varepsilon + \lambda \delta) \mathbb{E}_{\xi \sim P} \log \left(\mathbb{E}_{\zeta \sim \pi_0(\cdot | \xi)} e^{\frac{f(\zeta) - \lambda c(\xi, \zeta)}{\varepsilon + \lambda \delta}} \right). \quad (\text{C.3.36})$$

On the one hand, $g(\lambda^*)$ is upper bounded as

$$g(\lambda^*) \leq g(0) = \varepsilon \mathbb{E}_{\xi \sim P} \log \left(\mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi)} e^{\frac{f(\zeta)}{\varepsilon}} \right) \leq \sup_{\Xi} |f|. \quad (\text{C.3.37})$$

On the other hand, thanks to Jensen's inequality, $g(\lambda^*)$ is lower-bounded as

$$g(\lambda^*) \geq \lambda^* \rho + \mathbb{E}_{(\xi, \zeta) \sim \pi_0} [f(\zeta) - \lambda^* c(\xi, \zeta)] \geq \lambda^* (\rho - \mathbb{E}_{\pi_0} c) - \sup_{\Xi} |f|. \quad (\text{C.3.38})$$

Combining the two inequalities gives (C.3.35). $\blacksquare$

Lemma C.3.6 (Approximation bound for the Lagrangians). *Under the assumptions of Theorem C.3, consider*

$$F_{\rho}^{\varepsilon, \delta}(\lambda, f) = \sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi)} \mathbb{E}_{(\xi, \zeta) \sim \pi} [f(\zeta) - \lambda c(\xi, \zeta)] - (\varepsilon + \lambda \delta) \text{KL}(\pi | \pi_0), \quad (\text{C.3.39})$$

$$\text{and } G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(\lambda, f) = \sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi)} \mathbb{E}_{(\xi, \zeta) \sim \pi} [f(\zeta) - \lambda c(\xi, \zeta)] - \left(\frac{\varepsilon + \lambda \delta}{\sigma} \right) c(\xi, \zeta). \quad (\text{C.3.40})$$

Then we have, for a fixed $\Delta \in (0, d]$ and with $I_{\sigma}(\xi) := \int_{\Xi} e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} d\zeta$,

$$G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(\lambda, f) \leq F_{\rho}^{\varepsilon, \delta}(\lambda, f) + (\text{Lip}(f) + \lambda \text{Lip}(c))\Delta + (\varepsilon + \lambda \delta) \left(\frac{\Delta^p}{\sigma} - \log(V\Delta^d) + \mathbb{E}_{\xi \sim P} \log I_{\sigma}(\xi) \right). \quad (\text{C.3.41})$$

Proof. We start by reformulating $G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(\lambda, f)$. By continuity and compactness, the function $\zeta \mapsto f(\zeta) - (\lambda + (\varepsilon + \lambda \delta)/\sigma)c(\xi, \zeta)$ has a maximizer on Ξ for every ξ . By (Rockafellar and Wets, 1998, Thm. 14.37), we get that there exists a measurable map $\zeta^*: \Xi \rightarrow \Xi$ such that $\zeta^*(\xi) \in \arg \max_{\zeta \in \Xi} f(\zeta) - (\lambda + (\varepsilon + \lambda \delta)/\sigma)c(\xi, \zeta)$ for any $\xi \in \Xi$. Then,

$$\pi^*(d\xi, d\zeta) := P(d\xi) \delta_{\zeta^*(\xi)}(d\zeta) \quad (\text{C.3.42})$$

is an optimal solution, and therefore

$$G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(\lambda, f) = \mathbb{E}_{\pi^*} F_{\lambda} - \frac{\varepsilon + \lambda \delta}{\sigma} \mathbb{E}_{\pi^*} c. \quad (\text{C.3.43})$$

Now define $\overline{\mathbf{B}}^{\Delta}(\xi) := \overline{\mathbf{B}}(\zeta^*(\xi), \Delta)$ and $\pi^{\Delta} \in \mathcal{P}_P(\Xi \times \Xi)$ such that

$$\pi^{\Delta} \propto \mathbb{1}_{\zeta \in \overline{\mathbf{B}}^{\Delta}(\xi)} \pi_0(d\xi, d\zeta). \quad (\text{C.3.44})$$

Note first that we have

$$\mathbb{E}_{\pi^*} F_{\lambda} - \mathbb{E}_{\pi^{\Delta}} F_{\lambda} = \mathbb{E}_{\xi \sim P} \mathbb{E}_{\zeta \sim \pi^{\Delta}(\cdot|\xi)} [F_{\lambda}(\xi, \zeta^*(\xi)) - F_{\lambda}(\xi, \zeta)] \leq (\text{Lip}(f) + \lambda \text{Lip}(c))\Delta, \quad (\text{C.3.45})$$

since F_λ is $(Lip(f) + \lambda Lip(c))$ -Lipschitz continuous and the support of $\pi^\Delta(\cdot|\xi)$ is $\overline{\mathbb{B}}^\Delta(\xi)$. Now, we proceed to bound $KL(\pi^\Delta|\pi_0)$ by first noticing that

$$KL(\pi^\Delta|\pi_0) = \mathbb{E}_{\xi \sim P} \mathbb{E}_{\zeta \sim \pi^\Delta(\cdot|\xi)} \log \left(\frac{d\pi^\Delta(\zeta|\xi)}{d\pi_0(\zeta|\xi)} \right) \quad (\text{C.3.46})$$

$$= -\mathbb{E}_{\xi \sim P} \log \left(\int_{\Xi \cap \overline{\mathbb{B}}^\Delta(\xi)} e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} d\zeta \right) + \mathbb{E}_{\xi \sim P} \log \left(\int_{\Xi} e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} d\zeta \right) \quad (\text{C.3.47})$$

We focus on lower-bounding $\int_{\Xi \cap \overline{\mathbb{B}}^\Delta(\xi)} e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} d\zeta$. First, note that by the triangular inequality (and since $p \geq 1$), for any $\xi, \zeta \in \Xi$,

$$\frac{c(\xi, \zeta)}{2^p} \leq \frac{c(\xi, \zeta^\star(\xi)) + c(\zeta^\star(\xi), \zeta)}{2}. \quad (\text{C.3.48})$$

If, in addition, ζ is in $\overline{\mathbb{B}}^\Delta(\xi)$, this bound becomes $\frac{c(\xi, \zeta)}{2^{p-1}} \leq c(\xi, \zeta^\star(\xi)) + \Delta^p$. Hence, we have

$$\int_{\Xi \cap \overline{\mathbb{B}}^\Delta(\xi)} e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} d\zeta \geq e^{-\frac{c(\xi, \zeta^\star(\xi)) + \Delta^p}{\sigma}} \text{vol}(\Xi \cap \overline{\mathbb{B}}^\Delta(\xi)) \geq e^{-\frac{c(\xi, \zeta^\star(\xi)) + \Delta^p}{\sigma}} V \Delta^d, \quad (\text{C.3.49})$$

where V is defined in (C.3.25). Plugging the above lower and upper bounds into (C.3.47) yields

$$KL(\pi^\Delta|\pi_0) \leq \frac{\mathbb{E}_{\xi \sim P} c(\xi, \zeta^\star(\xi)) + \Delta^p}{\sigma} - \log(V \Delta^d) + \mathbb{E}_{\xi \sim P} \log I_\sigma(\xi) \quad (\text{C.3.50})$$

$$= \frac{\mathbb{E}_{\pi^\star} c + \Delta^p}{\sigma} - \log(V \Delta^d) + \mathbb{E}_{\xi \sim P} \log I_\sigma(\xi). \quad (\text{C.3.51})$$

Finally, putting (C.3.43), (C.3.45), and (C.3.51) together gives,

$$G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(\lambda, f) = \mathbb{E}_{\pi^\star} F_\lambda - \frac{\varepsilon + \lambda \delta}{\sigma} \mathbb{E}_{\pi^\star} c \quad (\text{C.3.52})$$

$$= \mathbb{E}_{\pi^\Delta} F_\lambda - (\varepsilon + \lambda \delta) KL(\pi^\Delta|\pi_0) + (\mathbb{E}_{\pi^\star} F_\lambda - \mathbb{E}_{\pi^\Delta} F_\lambda) + \left((\varepsilon + \lambda \delta) KL(\pi^\Delta|\pi_0) - \frac{\varepsilon + \lambda \delta}{\sigma} \mathbb{E}_{\pi^\star} c \right) \quad (\text{C.3.53})$$

$$\leq F_\rho^{\varepsilon, \delta}(\lambda, f) + (Lip(f) + \lambda Lip(c))\Delta + (\varepsilon + \lambda \delta) \left(\frac{\Delta^p}{\sigma} - \log(V \Delta^d) + \mathbb{E}_{\xi \sim P} \log I_\sigma(\xi) \right) \quad (\text{C.3.54})$$

which is the claimed inequality. $\blacksquare$

We have now all the ingredients to establish the extended version of the approximation result.

Proof of Theorem C.3. First, notice that by Lemma C.3.4, we have that

$$F_{\rho}^{0,0} \left(f, P \right) - \frac{\varepsilon \rho}{\sigma + \delta} = G_{\rho}^{0,0} \left(f \right) - \frac{\varepsilon \rho}{\sigma + \delta} \leq G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(f). \quad (\text{C.3.55})$$

Thus, using the bound at λ fixed, given by Lemma C.3.6 and the upper-bound (C.3.35), we get

$$G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(f) \leq \inf_{0 \leq \lambda \leq \bar{\lambda}} G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(\lambda, f) \quad (\text{C.3.56})$$

$$\leq \inf_{0 \leq \lambda \leq \bar{\lambda}} F_{\rho}^{\varepsilon, \delta}(\lambda, f) + (Lip(f) + \lambda Lip(c))\Delta + (\varepsilon + \lambda\delta) \left(\frac{\Delta^p}{\sigma} - \log(V\Delta^d) + \mathbb{E}_{\xi \sim P} \log I_{\sigma}(\xi) \right) \quad (\text{C.3.57})$$

$$\leq F_{\rho}^{\varepsilon, \delta}(f) + (Lip(f) + \bar{\lambda} Lip(c))\Delta + (\varepsilon + \bar{\lambda}\delta) \left(\frac{\Delta^p}{\sigma} - \log(V\Delta^d) + \mathbb{E}_{\xi \sim P} \log I_{\sigma}(\xi) \right). \quad (\text{C.3.58})$$

Minimizing the above bound over $\Delta > 0$, we get that the optimal value $\Delta^{\star}$ is of the form $\Delta^{\star} = \frac{(\varepsilon + \bar{\lambda}\delta)d}{L} + \mathcal{O}((\varepsilon + \bar{\lambda}\delta)^{p+1})$, with $L = Lip(f) + \bar{\lambda} Lip(c)$, as introduced in the statement of the theorem. As a consequence, we set $\Delta = \frac{(\varepsilon + \bar{\lambda}\delta)d}{L}$ in the bound above, which becomes

$$G_{\rho}^{\frac{\varepsilon}{\sigma}, \frac{\delta}{\sigma}}(f) \leq F_{\rho}^{\varepsilon, \delta}(f) + (\varepsilon + \bar{\lambda}\delta) \left(d + d \log \left(\frac{L}{(\varepsilon + \bar{\lambda}\delta)d} \right) + \mathbb{E}_{\xi \sim P} \log I_{\sigma}(\xi) + \log \frac{1}{V} + \frac{1}{\sigma} \left(\frac{(\varepsilon + \bar{\lambda}\delta)d}{L} \right)^p \right). \quad (\text{C.3.59})$$

There is only left to bound the term in $\mathbb{E}_{\xi \sim P} \log I_{\sigma}(\xi)$. On one hand, we have $e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} \leq 1$ so that

$$\int_{\Xi} e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} d\zeta \leq \text{vol}(\Xi).$$

On the other hand, we also have (using the change of variable $\zeta' = \sigma^{-\frac{1}{p}}(\zeta - \xi)$ to get I_{σ})

$$\int_{\Xi} e^{-\frac{c(\xi, \zeta)}{2^{p-1}\sigma}} d\zeta \leq \int_{\mathbb{R}^d} e^{-\frac{\|\xi - \zeta\|^p}{2^{p-1}\sigma}} d\zeta = I_{\sigma}$$

This makes the constant C appear in the bound and thus ends the proof. $\blacksquare$

We finish by explaining how the main theorem, [Theorem C.2](#), stems from [Theorem C.3](#). On the LHS of the inequality in [Theorem C.3](#), the unregularized objective has radius $\frac{\rho}{1+\delta/\sigma}$, instead of simply ρ in $F_{\rho}^{0,0}(f, P) = \text{val}(\text{WDRO})$. Thus, we compare in the next lemma the optimal values for these two parameters.

Lemma C.3.7 (Comparing optimal values). *Under the assumptions of [Theorem C.3](#),*

$$F_{\rho}^{0,0}(f, P) \leq F_{\frac{\rho}{1+\delta/\sigma}}^{0,0}(f, P) + \mathcal{O}(\delta). \quad (\text{C.3.60})$$

Proof. We first apply ([Santambrogio, 2015](#), Theorem 5.27) to get a constant-speed geodesic for the p -Wasserstein distance connecting P and Q , which is $W(P, Q)^{\frac{1}{p}}$ with our notation. This means that there exists a family of probability distributions $(Q_t)_{t \in [0,1]}$ such that $Q_0 = Q$, $Q_1 = P$ and, for any $t \in [0, 1]$,

$$W(P, Q_t)^{\frac{1}{p}} = (1-t)W(P, Q)^{\frac{1}{p}} \quad \text{and} \quad W(Q_t, Q)^{\frac{1}{p}} = tW(P, Q)^{\frac{1}{p}}. \quad (\text{C.3.61})$$

We apply these equations with Q such that $W(P, Q) \leq \rho$ and $t = 1 - (1 + \delta/\sigma)^{-\frac{1}{p}}$ to obtain

$$W(P, Q_t) \leq \frac{\rho}{1 + \delta/\sigma} \quad \text{and} \quad W(Q, Q_t) \leq t^p \rho = \mathcal{O}(\delta). \quad (\text{C.3.62})$$

Note that the first inequality above yields

$$\mathbb{E}_{Q_t} f \leq \sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi): \mathbb{E}_{\pi} c \leq \frac{\rho}{1+\delta/\sigma}} \mathbb{E}_{\pi_2} f = F_{\frac{\rho}{1+\delta/\sigma}}^{0,0}(f, P), \quad (\text{C.3.63})$$

We now use the Kantorovich-Rubinstein inequality (e.g., [Villani \(2003, Thm. 1.14\)](#)) to write

$$\mathbb{E}_Q f - \mathbb{E}_{Q_t} f \leq W_1(Q, Q_t) \text{Lip}(f) \leq W(Q, Q_t)^{\frac{1}{p}} \text{Lip}(f), \quad (\text{C.3.64})$$

where for the second inequality we used that the p -Wasserstein distance is always greater than or equal to the 1-Wasserstein distance (e.g., [Villani \(2003, §7.1.2\)](#)). Together with (C.3.63), this yields:

$$\mathbb{E}_Q f \leq \mathbb{E}_{Q_t} f + \mathcal{O}(\delta) \leq F_{\frac{\rho}{1+\delta/\sigma}}^{0,0}(f, P) + \mathcal{O}(\delta). \quad (\text{C.3.65})$$

Taking the supremum over all Q such that $W(P, Q) \leq \rho$ allows us to conclude. ■

With the help of the previous lemma, the proof of [Theorem C.2](#) comes easily from [Theorem C.3](#).

Proof of Theorem C.2. We start with checking that the fact that Ξ is a compact convex body (condition (iv) in [Theorem C.2](#)) implies that $V > 0$ (condition (iv) in [Theorem C.3](#)). Introduce, for $\xi \in \Xi$, the function

$$\nu_\xi : \Delta \mapsto \frac{\text{vol}(\Xi \cap \overline{\mathbb{B}}(\xi, \Delta))}{\Delta^d} = \text{vol}\left(\frac{1}{\Delta}(\Xi - \xi) \cap \overline{\mathbb{B}}(0, 1)\right).$$

Since Ξ is convex, we easily get that ν_ξ is non-increasing. Thus we can lower-bound the constant V as follows, using $\text{diam}(\Xi) := \sup_{\xi, \zeta \in \Xi} \|\xi - \zeta\|$ the diameter of Ξ .

$$V = \inf_{\xi \in \Xi, 0 < \Delta \leq d} \nu_\xi(\Delta) \geq \inf_{\xi \in \Xi, 0 < \Delta \leq \max(d, \text{diam}(\Xi))} \nu_\xi(\Delta) \quad (\text{C.3.66})$$

$$= \inf_{\xi \in \Xi} \frac{\text{vol}(\Xi \cap \overline{\mathbb{B}}(\xi, \max(d, \text{diam}(\Xi))))}{(\max(d, \text{diam}(\Xi)))^d} \geq \frac{\text{vol}(\Xi)}{(\max(d, \text{diam}(\Xi)))^d} > 0. \quad (\text{C.3.67})$$

So we get $V > 0$ which is condition (iv) in [Theorem C.3](#). Thus we can apply [Theorem C.3](#) and [Lemma C.3.7](#). Noting that $F_{\rho}^{\varepsilon, \delta}(f) = \text{val}(\text{E-WDRO})$ and $F_{\rho}^{0,0}(f, P) = \text{val}(\text{WDRO})$, and combining the obtained bound with [Lemma C.3.7](#) gives the result. ■

C.4 CONCLUSION, PERSPECTIVES

Inspired by the success of regularization in OT, we proposed and studied a regularization scheme for WDRO problems. We derived the expression of the dual objective function in the general case as well as a refined one in the particular setting of the entropic regularization. In addition, we showed that the difference between the original WDRO problem and the entropic one is properly controlled by the regularization parameters.

Since regularization in OT has shown attractive computational advantages and statistical benefits, an exciting research direction is to investigate whether similar gains hold for regularization in WDRO. It is also worth studying possible extensions of the results of this paper, including the approximation errors with general regularizations, beyond the entropic case.

C.5 DUALITY UNDER WEAKER ASSUMPTIONS

In [Theorem C.1](#), we derive the expression of the dual of the regularized WDRO problem, with arbitrary regularizations. In this situation, the primal is not completely explicit, and then compactness and continuity are crucial, as we leverage the duality between continuous functions and measures on a compact space.

In this appendix, we provide a generalization of the duality result that does away with the compactness and the continuity assumptions. This is achieved by carefully designing the spaces in which duality arguments are made. The general duality result was informally given in [Appendix C.2.3](#). Here, we provide the mathematical developments, as follows: [Appendix C.5.1](#) presents the setting; [Appendix C.5.2](#) provides the theorem; [Appendix C.5.3](#) states some basic results used in the the proof of [Appendix C.5.4](#); [Appendix C.5.5](#) finishes with two examples.

C.5.1 Setting: assumptions, recalls, and construction of the spaces

In the appendix, we consider the following set of blanket assumptions.

Assumption C.3.

- (i) $\Xi \subset \mathbb{R}^d$ is a closed set.
- (ii) $f: \Xi \rightarrow \mathbb{R}$ is measurable w.r.t. the Borel σ -algebra on Ξ ;
- (iii) $c: \Xi \times \Xi \rightarrow \mathbb{R}_+$ is measurable as well and, for all ξ in Ξ , $c(\xi, \xi) = 0$;
- (iv) There is some $\xi_0 \in \Xi$ such that the integrability condition $\mathbb{E}_{\xi \sim P}[c(\xi, \xi_0)] < +\infty$ holds.

For convenience, we denote by c_0 the function $(\xi, \zeta) \mapsto c(\xi, \xi_0)$. We work in this appendix with the probability space $(\Xi^2, \mathcal{F}^{\otimes 2}, \pi_0)$ where π_0 is a measure satisfying the following assumption.

Assumption C.4. $\pi_0 \in \mathcal{P}_P(\Xi \times \Xi)$ is a reference measure such that $1 + c_0 + c$ is integrable w.r.t. π_0 .

As a consequence of (ii), we consider $L^0(\pi_0)$ the set of measurable functions on Ξ^2 (where functions which agree almost everywhere w.r.t. π_0 are identified), equipped with the topology which metrizes the convergence in probability, see e.g., [Kalton et al. \(1984, Chap. 2, §2\)](#) or [Bogachev \(2007, 4.7.60\)](#). For simplicity, we identify $L^0(\pi_0)$, and more generally the spaces $L^p(\pi_0)$, as subsets of $\mathcal{M}(\Xi^2)$ through the map $g \mapsto g \, d\pi_0$. In turn, $\mathcal{M}(\Xi^2)$ is itself seen as a subspace of $(L^b(\Xi^2))^*$, the topological dual of the space of bounded (everywhere) measurable functions.

Recalls in modular spaces

The upcoming results are built over modular spaces [Musielak \(2006\)](#), i.e., subsets of $L^1(\pi_0)$ defined for some $H: L^0(\pi_0) \rightarrow \mathbb{R} \cup \{+\infty\}$ as

$$L^H := \{g \in L^1(\pi_0) : \exists \beta > 0, H(g/\beta) < +\infty\}. \quad (\text{C.5.1})$$

Under suitable assumptions, the following lemma states that L^H equipped with the norm $\|g\|_H := \inf\{\beta > 0 : H(g/\beta) \leq 1\}$ is a Banach space.

Lemma C.5.1. Consider $H : L^0(\pi_0) \rightarrow \mathbb{R}_+ \cup \{+\infty\}$ convex, symmetric, with $H(0) = 0$. If H is $L^0(\pi_0)$ -l.s.c. and $\{g \in L^0(\pi_0) : H(g) \leq 1\}$ is bounded in $L^0(\pi_0)$, then $(L^H, \|\cdot\|_H)$ is a Banach space.

Proof. The proof follows the second paragraph of the proof of Pennanen and Perkkiö (2022, Thm. 22), which consists in applying Pennanen and Perkkiö (2022, Rmk. 9). For this, one needs to show that the topology on L^H is no weaker than the $L^0(\pi_0)$ topology. First note that the $L^0(\pi_0)$ -l.s.c. of H implies the $L^0(\pi_0)$ -l.s.c. of $\|\cdot\|_{L^H}$. Indeed, take a sequence g_t of L^H which converges to g : for any $\alpha > \|g_t\|_{L^H}$, it holds that $H(g_t/\alpha) \leq 1$.

Now, take $\mathcal{U} \subset L^H$ an $L^0(\pi_0)$ -neighborhood of the origin. By the boundedness assumption, there is $\alpha > 0$ such that $\{g \in L^H : \|g\|_H \leq 1\} = \{g \in L^0(\pi_0) : H(g) \leq 1\} \subset \alpha \mathcal{U}$ so that $\mathcal{U} \cap L^H$ is a neighborhood of 0 in L^H . Applying Pennanen and Perkkiö (2022, Rmk. 9) yields the result. ■

We also consider a special subspace of L^H , named the Orlicz heart, denoted by $L_\heartsuit^H$ and defined by

$$L_\heartsuit^H := \{g \in L^H : \forall \beta > 0, H(g/\beta) < +\infty\}. \quad (\text{C.5.2})$$

Our motivation for introducing such spaces was guided by Agrawal and Horel (2021), which explores the link between ϕ -divergences and associated Orlicz spaces. Note though that our construction is more general and slightly different than theirs.

Construction of the function space for duality

We first introduce two modular spaces, L^{H_c} and L^{H_R} , used later to build our function space $\mathcal{X}$.

Lemma C.5.2. Define $H_c : L^0(\pi_0) \rightarrow \mathbb{R} \cup \{+\infty\}$ as

$$H_c := \iota \left\{ g \in L^0(\pi_0) : \left\| \frac{g}{1+c_0+c} \right\|_{L^\infty(\pi_0)} \leq 1 \right\}. \quad (\text{C.5.3})$$

Then $(L^{H_c}, \|\cdot\|_{H_c})$ is a Banach space.

Proof. We check that H_c satisfies the assumptions of Lemma C.5.1, and in particular that it is $L^0(\pi_0)$ -l.s.c. and that $\{g \in L^0(\pi_0) : H_c(g) \leq 1\}$ is bounded in probability, which comes down to showing that $\left\{ g \in L^0(\pi_0) : \|g/(1+c_0+c)\|_{L^\infty(\pi_0)} \leq 1 \right\}$ is closed in $L^0(\pi_0)$ and bounded in $L^0(\pi_0)$.

To show that it is closed, consider a sequence of functions g_t for $t = 1, 2, \dots$ belonging to this set which converges in probability to some $g \in L^0(\pi_0)$. Then, for any $\eta > 0$,

$$\pi_0 \left(\frac{|g|}{1+c_0+c} > 1+\eta \right) \leq \pi_0 \left(\frac{|g-g_t|}{1+c_0+c} > \eta \right) \pi_0 \left(\frac{|g_t|}{1+c_0+c} > 1 \right) \quad (\text{C.5.4})$$

$$\leq \pi_0(|g-g_t| > \eta) \xrightarrow{t \rightarrow +\infty} 0. \quad (\text{C.5.5})$$

Hence, by continuity of probability,

$$\pi_0 \left(\frac{|g|}{1+c_0+c} > 1 \right) = \lim_{\eta \rightarrow 0} \pi_0 \left(\frac{|g|}{1+c_0+c} > 1+\eta \right) = 0, \quad (\text{C.5.6})$$

which shows that $\|g/(1+c_0+c)\|_{L^\infty(\Xi)} \leq 1$.

We now show that $\{g \in L^0(\pi_0) : \|g/(1+c_0+c)\|_{L^\infty(\Xi)} \leq 1\}$ is bounded in $L^0(\pi_0)$ i.e.,

$$\forall \eta > 0, \exists M > 0, \forall g \in L^0(\pi_0) \quad \text{s.t.} \quad H(g) \leq 1, \pi_0(|g| \geq M) \leq \eta. \quad (\text{C.5.7})$$

To do so, let us take $g \in L^0(\pi_0)$ such that $\|g/(1+c_0+c)\|_{L^\infty(\pi_0)} \leq 1$. Then, for any $M > 0$, by Markov's inequality

$$\pi_0(|g| \geq M) \leq \frac{\mathbb{E}_{\pi_0}[|g|]}{M} \leq \frac{\mathbb{E}_{\pi_0}[1+c_0+c]}{M}, \quad (\text{C.5.8})$$

and by assumption, $\mathbb{E}_{\pi_0}[1+c_0+c]$ is finite, so that the RHS is arbitrarily small as M grows. ■

Now, let us define a modular space, starting from a “regularization” R (a more formal link will be made in [Appendix C.5.5](#)).

Lemma C.5.3. *Consider $R : (L^b(\Xi^2))^* \rightarrow \mathbb{R}_+$ convex such that*

1. *dom R is included in the cone of non-negative linear forms;*
2. *$\inf R = 0 = R(\pi_R)$ for some $\pi_R \in L^\infty(\pi_0)$ such that both $(1+\alpha)\pi_R$ and $\pi_R + \alpha$ are in dom R for some $\alpha > 0$;*

and define $H_R : L^0(\pi_0) \rightarrow \mathbb{R} \cup \{+\infty\}$ by

$$H_R(g) := \sup_{\pi \in L^0(\pi_0)} \int_{\Xi^2} |g| \pi d\pi_0 - R(\pi + \pi_R). \quad (\text{C.5.9})$$

Then $(L^{H_R}, \|\cdot\|_{H_R})$ is a Banach space. As a consequence, $(L_{\heartsuit}^{H_R}, \|\cdot\|_{H_R})$ is a Banach space too.

Proof. Note first that $H_R(g) = \sup_{\pi \in L^0(\pi_0): \pi \geq 0} \int_{\Xi^2} |g| \pi d\pi_0 - \int_{\Xi^2} |g| \pi_R d\pi_0 - R(\pi)$ so that it is well-defined for any $g \in L^1(\pi_0)$ since the integral $\int_{\Xi^2} |g| \pi d\pi_0$ always has a value in $\mathbb{R} \cup \{+\infty\}$ and because $g\pi_R$ is still in $L^1(\pi_0)$. Moreover H_R is convex as a supremum of convex functions.

Let us show that H_R satisfies the conditions of [Lemma C.5.1](#). First, H_R is non-negative since

$$H_R(g) \geq \langle 0 | g \rangle - R(0 + \pi_R) = 0 \quad \text{for } g \in L^0(\pi_0). \quad (\text{C.5.10})$$

Then, H_R is $L^0(\pi_0)$ -l.s.c. since the functions $g \mapsto \int |g|(\pi - \pi_R) d\pi_0 - R(\pi)$ are $L^0(\pi_0)$ -l.s.c. for any $\pi \in L^0(\pi_0)$ non-negative by Fatou's lemma. Finally, it remains to check the boundedness condition, i.e., that $\{g \in L^0(\pi_0) : H_R(g) \leq 1\}$ is bounded in $L^0(\pi_0)$. But, since $\pi_R + \alpha \in \text{dom } R$ for some $\alpha > 0$, for any $g \in L^0(\pi_0)$, $H_R(g) \geq \alpha \|g\|_{L^1(\pi_0)}$ so that $\{g \in L^0(\pi_0) : H_R(g) \leq 1\}$ is included in a $L^1(\pi_0)$ -ball and is *a fortiori* bounded in probability by Markov inequality.

For $L_{\heartsuit}^{H_R}$ to be Banach space too, it suffices to check that it is closed in L^{H_R} , following the arguments of the proof of [Pennanen and Perkkiö \(2022, Thm. 22\)](#). Indeed, take a sequence $(g_t)_t$ from $L_{\heartsuit}^{H_R}$ that converges to $g \in L^H$ for the norm $\|\cdot\|_{L^{H_R}}$. For any $\beta > 0$, $H_R((g - g_t)/\beta) \leq 1$ for t large enough so the convexity of H_R gives us that,

$$H_R(g/(2\beta)) \leq \frac{1}{2} H_R(g_t/\beta) + \frac{1}{2} < +\infty, \quad (\text{C.5.11})$$

so that $g \in L_{\heartsuit}^{H_R}$, and this ends the proof. ■

As a consequence of these two lemmas, we can build the Banach space $\mathcal{X}$ that will be used in the duality proofs, see [Lunardi \(2009, p. ix\)](#) for instance.

Corollary C.5.4. Define $\mathcal{X} = L^{H_c} + L_{\heartsuit}^{H_R}$ with the norm

$$\|g\|_{\mathcal{X}} = \inf \left\{ \|g_c\|_{L^{H_c}} + \|g_R\|_{L^{H_R}} : g_c \in L^{H_c}, g_R \in L_{\heartsuit}^{H_R} \text{ s.t. } g_c + g_R = g \right\}. \quad (\text{C.5.12})$$

Then $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is a Banach space, whose dual can be identified with

$$\mathcal{X}^* = (L^{H_c})^* \cap (L_{\heartsuit}^{H_R})^*. \quad (\text{C.5.13})$$

Moreover, we will need the following criterion under which a given measure belongs to $\mathcal{X}$.

Lemma C.5.5. Under the assumptions of [Lemma C.5.3](#), for $\pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0)$, if both $\mathbb{E}_{\pi}[c]$ and $R(\pi)$ are finite, then π belongs to $\mathcal{X}^*$.

Proof. We first show that π is in $(L^{H_c})^*$. To do so, we use that by [Pennanen and Perkkiö \(2022, Thm. 22\)](#), L^{H_c} is a subspace of $(L^{H_c})^*$ where H_c^* is defined for all $\pi \in (L^{H_c})^*$ as $H_c^*(\pi) = \sup_{g \in L^{\infty}(\pi_0)} \langle \pi, g \rangle - H_c(g)$. By definition of H_c , we have that $H_c^*(\pi) = \langle \pi, 1 \rangle + c_0 + c = 1 + \mathbb{E}_{\xi \sim P}[c(\xi, \xi_0)] + \mathbb{E}_{\pi}[c]$, which is finite by assumption. We thus have that $\pi \in (L^{H_c})^*$.

Next, we show that π defines a continuous linear form on $(L_{\heartsuit}^{H_R}, \|\cdot\|_{L^{H_R}})$. To do so, let us take $g \in L_{\heartsuit}^{H_R}$ so that by definition we can take $\beta > 0$ so that $H_R(g/\beta) < +\infty$. By the assumption of [Lemma C.5.3](#), there is some $\alpha > 0$ such that $(1 + \alpha)\pi_{\varepsilon} \in \text{dom } R$ so that one can choose $\lambda \in (0, 1)$ satisfying $(1 + \alpha)(1 - \lambda) = 1$. Then, by definition of H_R and convexity of R , one has that for all $\pi \in L^0(\pi_0)$, the following inequality holds

$$\frac{\lambda}{\beta} \int |g| \pi d\pi_0 \leq H_R(g/\beta) + \lambda R(\pi) + (1 - \lambda)R((1 + \alpha)\pi_R). \quad (\text{C.5.14})$$

Since the RHS is finite from the assumptions as soon as $R(\pi)$ is finite, we have that $\int g \pi d\pi_0$ is well-defined for $g \in L_{\heartsuit}^{H_R}$. All that is left to show is that this linear form is indeed continuous. Take $g_t \in L_{\heartsuit}^{H_R}$ which goes to zero as t goes to $+\infty$. This means that there exists a sequence $\beta_t > 0$ which goes to zero such that $H_R(g_t/\beta_t) \leq 1$ for all $t = 1, 2, \dots$. Applying (C.5.14) with $g \leftarrow g_t, \beta \leftarrow \beta_t$, and π such that $R(\pi)$ is finite yields

$$\frac{\lambda}{\beta_t} \int |g_t| \pi d\pi_0 \leq 1 + \lambda R(\pi) + (1 - \lambda)R((1 + \alpha)\pi_R) < +\infty \quad (\text{C.5.15})$$

which implies that $\int |g_t| \pi d\pi_0$ goes to 0 as t goes to $+\infty$. ■

c.5.2 Statement of the theorem

We can now state the formal version of [Theorem C.4](#).

Theorem C.6. Take $S := \frac{\delta}{\varepsilon}R$. Let [Assumption C.3](#) hold and assume that

1. f satisfies the growth condition :

$$\exists g \in L^{\infty}(\pi_0), \forall \beta > 0, (\xi, \zeta) \mapsto \beta(f(\zeta) - (1 + c(\xi, \xi_0) + c(\xi, \zeta))g(\xi, \zeta)) \in \text{dom } H_R, \quad (\text{C.5.16})$$

or equivalently, that the function $\tilde{f}$, defined as $\tilde{f} : (\xi, \zeta) \mapsto f(\zeta)$, lies in $\mathcal{X}$;

2. R convex satisfies the assumptions of [Lemma C.5.3](#) and $R|_{\mathcal{X}^*}$ is $\sigma(\mathcal{X}^*, \mathcal{X})$ -l.s.c.;
3. $\pi_0 \in \mathcal{P}_P(\Xi \times \Xi)$ satisfies [Assumption C.4](#);
4. there exists $\pi_S \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0)$ such that $\mathbb{E}_{\pi_S}[c] + S(\pi_S) < \rho$;
5. one of the two following regularity conditions hold:
 - a) R is $\sigma(\mathcal{X}^*, \mathcal{C}(\Xi^2) \cap L^{H_c})$ -u.s.c. on $\mathcal{P}_P(\Xi \times \Xi)$, f is l.s.c., $f_- \in L^{H_c}$, $\overline{\text{int}} \Xi = \Xi$ and for all $\xi \in \Xi$ and $\pi_0(\cdot|\xi)$ has a positive density w.r.t. the Lebesgue measure on Ξ .
 - b) $\text{dom } R \subset L^0(\pi_0)$ and, for any non-increasing non-negative sequence $g_t \in \text{dom } H_R$ such that $g_t \rightarrow 0$ as $t \rightarrow \infty$, we have $H_R(g_t) \rightarrow 0$.

Then, the problem

$$\sup \{ \mathbb{E}_{\pi_2}[f] - R(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi), \mathbb{E}_{\pi}[c] + S(\pi) \leq \rho \} \quad (\text{C.5.17})$$

coincides with its dual

$$\inf \left\{ \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\text{ess sup}_{\pi_0(\cdot|\xi)} \{ f - \lambda c(\xi, \cdot) - g(\xi, \cdot) \} \right] + (R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*})_*(g) : g \in \mathcal{X}, \lambda \geq 0 \right\}. \quad (\text{C.5.18})$$

We state the theorem and provide its proof in the case of when R and S are proportional. The case with different R and S would follow that the same lines, at the price of slightly more complicated assumptions and a slightly different treatment of subcases in the proof. To avoid these extra-technicalities, we stick here with the proportional case, which already presents all the intrinsic reasoning and the mathematical tools. Before going to the proof of this result in [Appendix C.5.4](#), we introduce in the next section some general duality lemmas, tailored for our context.

C.5.3 Basic duality lemmas

Lemma C.5.7. Consider $\mathcal{X}$ a Banach space and $g : \mathcal{X}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ convex, proper and weakly- $\star$ l.s.c.. Then $(g_*)^* = g$.

Proof. Equipped with its weak- $\star$ topology, $\mathcal{X}^*$ is a locally convex topological vector space whose topological dual is $\mathcal{X}$, see [Rudin \(1991, Thm. 3.10, §3.14\)](#). As a consequence, the usual theorem of the biconjugate (e.g., [Zalinescu \(2002, Thm. 2.3.3\)](#)) ensures that $(g_*)^* = g$. ■

Lemma C.5.8. Given $\mathcal{X}$ a Banach space, $g, h, s, l : \mathcal{X}^* \rightarrow \mathbb{R} \cup \{+\infty\}$ convex, proper and weakly- $\star$ l.s.c., consider the two problems,

$$\min_{x \in \mathcal{X}, \lambda \geq 0} (g + \lambda s)_*(x) + (h + \lambda l)_*(-x) \quad (\text{P})$$

$$\max_{x^* \in \mathcal{X}^*} -g(x^*) - h(x^*) \text{ s.t. } s(x^*) + l(x^*) \leq 0. \quad (\text{D})$$

If the following qualification condition holds

$$\{x \in \mathcal{X} : \exists \lambda \geq 0, x \in \text{dom}(g + \lambda s)_* + \text{dom}(h + \lambda l)_*\} = \mathcal{X} \quad (\text{C.5.19})$$

then, $(\text{P}) = (\text{D})$.

Proof. Let us define the perturbation function $\Phi : (\mathcal{X} \times \mathbb{R}) \times \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ by

$$\Phi((x, \lambda), x') = (g + \lambda s)_*(x) + (h + \lambda l)_*(x' - x) + \iota_{\mathbb{R}_+}(\lambda). \quad (\text{C.5.20})$$

Φ is l.s.c. and convex since both $(x, \lambda), x' \mapsto (g + \lambda s)_*(x)$ and $(x, \lambda), x' \mapsto (h + \lambda l)_*(x' - x)$ are jointly convex and l.s.c. as supremums of continuous linear functions. Let us now compute the conjugate of Φ , for $((x^*, \lambda^*), y) \in (\mathcal{X}^* \times \mathbb{R}) \times \mathcal{X}^*$

$$\Phi^*((x^*, \lambda^*), y) = \sup_{x, x' \in \mathcal{X}, \lambda \geq 0} \{ \langle x^*, x \rangle + \langle y, x' \rangle + \lambda^* \lambda - (g + \lambda s)_*(x) - (h + \lambda l)_*(x' - x) \} \quad (\text{C.5.21})$$

$$= \sup_{x, x' \in \mathcal{X}, \lambda \geq 0} \{ \langle x^*, x \rangle + \langle y, x' + x \rangle + \lambda^* \lambda - (g + \lambda s)_*(x) - (h + \lambda l)_*(x') \} \quad (\text{C.5.22})$$

$$= \sup_{\lambda \geq 0} \left\{ \lambda^* \lambda + (g + \lambda s)(x^* + y) + \sup_{x \in \mathcal{X}} \{ \langle x^*, x \rangle - (h + \lambda l)_*(x') \} \right\} \quad (\text{C.5.23})$$

$$= \sup_{\lambda \geq 0} \{ \lambda^* \lambda + (g + \lambda s)(x^* + y) + (h + \lambda l)(y) \} \quad (\text{C.5.24})$$

$$= g(x^* + y) + h(y) + \iota_{\lambda^* + s(x^* + y) + l(y) \leq 0}, \quad (\text{C.5.25})$$

where we applied [Lemma C.5.7](#) to $g + \lambda s$ and $h + \lambda l$. Now, note that (P) is equal to

$$\min_{(x, \lambda) \in (\mathcal{X}, \mathbb{R})} \Phi((x, \lambda), 0) \quad \text{and (D) to} \quad \max_{x^* \in \mathcal{X}^*} -\Phi^*(0, x^*). \quad (\text{C.5.26})$$

Now we leverage the duality theorem [Zalinescu \(2002, Thm. 2.7.1\)](#) with qualification condition (vii), see [Zalinescu \(2002, p. 15\)](#) for the relevant definition, to show that the values of those problems are equal. This qualification condition requires that the set

$$\text{cone}\{x' \in \mathcal{X} : \exists (x, \lambda) \in \mathcal{X} \times \mathbb{R}, ((x, \lambda), x') \in \text{dom } \Phi\} \quad (\text{C.5.27})$$

be a closed linear subset of $\mathcal{X}$. But this set can be rewritten as

$$\text{cone}\{x' \in \mathcal{X} : \exists \lambda \geq 0, x' \in \text{dom}(g + \lambda s)_* + \text{dom}(h + \lambda l)_*\} \quad (\text{C.5.28})$$

and our qualification assumption ensures that it is equal to the whole space $\mathcal{X}$. ■

Lemma C.5.9. Consider a vector space $\mathcal{Y}$, $g : \mathcal{Y} \rightarrow \{-\infty\} \cup \mathbb{R}$ concave, $s : \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ convex and assume that there exists $y_s \in \text{dom } g$ such that $s(y_s) < 0$. Then,

$$\sup\{g(y) : y \in \mathcal{Y}, s(y) \leq 0\} = \sup\{g(y) : y \in \mathcal{Y}, s(y) < 0\}. \quad (\text{C.5.29})$$

Proof. The inequality ($\geq$) always holds. To prove the reverse, take $y \in \text{dom } g$ such that $s(y) \leq 0$, consider $y_t := ty_s + (1 - t)y$ for $t \in [0, 1]$. By convexity of s and definition of y_s , we have that $s(y_t) < 0$ for all $t \in (0, 1]$. Moreover, since y_s is in $\text{dom } g$ and g is concave, it holds that

$$g(y) \leq \liminf_{t \rightarrow 0} \frac{g(y_t) - tg(y_s)}{1 - t} = \liminf_{t \rightarrow 0} g(y_t) \leq \sup\{g(y) : y \in \mathcal{Y}, s(y) < 0\}, \quad (\text{C.5.30})$$

which shows the required inequality. ■

c.5.4 Proof of Theorem C.6

The proof of Theorem C.6 is divided into two main lemmas, Lemma C.5.11 and Lemma C.5.13, which, put together, exactly give Theorem C.6.

Lemma C.5.10. Assume that $\mathcal{X} \subset L^1(\pi_0)$ is a Banach space then the support function of the set $\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$, seen as a subspace of $\mathcal{X}^*$, is given for all $g \in \mathcal{X}$ by

$$\sigma_{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}(g) := \sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*} \langle \pi, g \rangle = \mathbb{E}_{\xi \sim P} \left[\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi, \cdot) \right] \quad (\text{C.5.31})$$

Remark 11. We observe that $\mathbb{E}_{\xi \sim P} \left[\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi, \cdot) \right]$ is well-defined for $g \in \mathcal{X}$. First, the map $\xi \mapsto \operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi, \cdot)$ is measurable since, for any $\alpha \in \mathbb{R}$, by definition of the essential supremum,

$$\left\{ \xi \in \Xi : \operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi, \cdot) > \alpha \right\} = \left\{ \xi \in \Xi : \pi_0(\{\zeta \in \Xi : g(\xi, \zeta) > \alpha\} | \xi) > 0 \right\} \quad (\text{C.5.32})$$

$$= \left\{ \xi \in \Xi : \int_{\Xi} \mathbb{1}_{\{(\xi, \zeta) \in \Xi^2 : g(\xi, \zeta) > \alpha\}}(\xi, \zeta) d\pi_0(\zeta | \xi) > 0 \right\}, \quad (\text{C.5.33})$$

where the resulting integral is a measurable function of ξ , see e.g., [Kallenberg \(2017, Lem. 3.2 \(i\)\)](#). Then, if $g \in \mathcal{X}$, then for π_0 -almost every (ξ, ζ) , $\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi, \cdot) \geq g(\xi, \zeta)$ so that by integrating w.r.t. π_0 , one gets that $\mathbb{E}_{\xi \sim P} \left[\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi, \cdot) \right] \geq \mathbb{E}_{\pi_0}[g] > -\infty$ since $g \in \mathcal{X} \subset L^1(\pi_0)$.

Proof. Fix some $g \in \mathcal{X}$. First, let us show the inequality $(\leq)$. To do so, take $\pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$. By disintegration of measure, it holds that

$$\mathbb{E}_{\pi}[g] = \mathbb{E}_{\xi \sim P} \left[\mathbb{E}_{\zeta \sim \pi(\cdot|\xi)}[g(\xi, \zeta)] \right] \leq \mathbb{E}_{\xi \sim P} \left[\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi, \cdot) \right], \quad (\text{C.5.34})$$

since $\pi(\cdot|\xi)$ is a probability distribution which is absolutely continuous w.r.t. $\pi_0(\cdot|\xi)$ for any $\xi \in \Xi$.

To prove the reverse inequality, we distinguish three cases. For convenience, denote by h the measurable function $\xi \mapsto \operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} g(\xi, \cdot)$. Furthermore, note that the fact that $\mathcal{X}$ is a subset of $L^1(\pi_0)$ implies that $\mathcal{X}^*$ contains $L^\infty(\pi_0)$, so the indicator functions of measurable sets, in particular.

1. If $(P\text{-a.s. } g(\xi, \cdot) \in L^\infty(\pi_0(\cdot|\xi)))$ is false, i.e., if h is not finite $P\text{-a.s.}$. This means that, for any $n \geq 1$, the sets $A_n(\xi) := \{\zeta \in \Xi : g(\xi, \zeta) \geq n\}$ satisfy $P\text{-a.s.}, \pi_0(A_n(\xi)|\xi) > 0$. Therefore, one can define $\pi_n \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$ as the measurable map $(\xi, \zeta) \mapsto \frac{\mathbb{1}_{A_n(\xi)}(\zeta)}{\pi_0(A_n(\xi)|\xi)}$ and it satisfies $\langle \pi_n, g \rangle \geq n$. This shows that

$$\sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*} \langle \pi, g \rangle = +\infty = \mathbb{E}_{\xi \sim P}[h(\xi)]. \quad (\text{C.5.35})$$

2. If $P\text{-a.s. } g(\xi, \cdot) \in L^\infty(\pi_0(\cdot|\xi))$, i.e., if h is finite $P\text{-a.s.}$ and $h \in L^1(P)$. As a consequence of the definition of h , for any $n \geq 1$, the sets $A_n(\xi) := \{\zeta \in \Xi :$

$g(\xi, \zeta) \geq h(\xi) - 1/n$ satisfy P a.s., $\pi_0(A_n(\xi)|\xi) > 0$. Again, one defines $\pi_n \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$ as the map $(\xi, \zeta) \mapsto \frac{\mathbb{1}_{A_n(\xi)}(\zeta)}{\pi_0(A_n(\xi)|\xi)}$ so that it satisfies $\langle \pi_n, g \rangle \geq \mathbb{E}_P[h] - 1/n$ for all $n \geq 1$ and thus

$$\sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*} \langle \pi, g \rangle \geq \mathbb{E}_{\xi \sim P}[h(\xi)]. \quad (\text{C.5.36})$$

3. Finally, if P-a.s. $g(\xi, \cdot) \in L^\infty(\pi_0(\cdot|\xi))$, i.e., if h is finite P-a.s. but $h \notin L^1(P)$. By Remark 11, this means that $\mathbb{E}_P[h] = +\infty$. By mimicking the proof of the previous point but with $h_n(\xi, \zeta) := h(\xi) \mathbb{1}_{h(\xi) \leq n} + g(\xi, \zeta) \mathbb{1}_{h(\xi) > n}$ instead of h , we obtain that, for any $n \geq 1$,

$$\sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*} \langle \pi, g \rangle \geq \mathbb{E}_{\xi \sim \pi_0}[h_n(\xi, \zeta)]. \quad (\text{C.5.37})$$

Now, by applying the monotone convergence theorem to $h_n - g$, which is non-negative π_0 -a.s., and with $h \in L^1(\pi_0)$ by assumption, we get that

$$\sup_{\pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*} \langle \pi, g \rangle \geq \mathbb{E}_{\xi \sim P}[h(\xi)] = +\infty, \quad (\text{C.5.38})$$

which concludes the proof. $\blacksquare$

Lemma C.5.11. *Under the assumptions of Theorem C.6, the problem*

$$\sup \left\{ \mathbb{E}_{\pi_2}[f] - R(\pi) : \pi \in \overline{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}, \langle \pi, c \rangle + S(\pi) \leq \rho \right\} \quad (\text{C.5.39})$$

coincides with its dual

$$\inf \left\{ \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\text{ess sup}_{\pi_0(\cdot|\xi)} \{f - \lambda c(\xi, \cdot) - g(\xi, \cdot)\} \right] + (R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*})_*(g) : g \in \mathcal{X}, \lambda \geq 0 \right\}. \quad (\text{C.5.40})$$

Proof. We apply Lemma C.5.8 with $\mathcal{X}$ and $g \leftarrow R|_{\mathcal{X}^*}$, $s \leftarrow S|_{\mathcal{X}^*}$ which are convex, proper and $\sigma(\mathcal{X}^*, \mathcal{X})$ -l.s.c.. Indeed, $\sigma(\mathcal{X}^*, L^{H_\epsilon})$ is weaker than $\sigma(\mathcal{X}^*, \mathcal{X})$, and $h \leftarrow \langle \cdot \rangle - \tilde{f} + \iota_{\overline{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}}$, $l \leftarrow \langle \cdot \rangle c - \rho$ are convex, proper and weakly- $\star$ l.s.c. by construction. For $\lambda \geq 0$, we have

$$\forall g \in \mathcal{X}, \quad (h + \lambda l)_*(g) = \lambda \rho + \sup_{\pi \in \overline{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}} \langle \pi, g \rangle - \langle \pi, \lambda c \rangle - \tilde{f} \quad (\text{C.5.41})$$

$$= \lambda \rho + \sigma_{\overline{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}}(\tilde{f} - \lambda c + g) \quad (\text{C.5.42})$$

$$= \lambda \rho + \sigma_{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}(\tilde{f} - \lambda c + g) \quad (\text{C.5.43})$$

$$= \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\text{ess sup}_{\pi_0(\cdot|\xi)} \{f - \lambda c(\xi, \cdot) - g(\xi, \cdot)\} \right] \quad (\text{C.5.44})$$

where the last equality comes from Lemma C.5.10. The qualification condition of Lemma C.5.8 writes,

$$\left\{ g \in \mathcal{X} : \exists \lambda \geq 0, g \in \text{dom}(R + \lambda S)_* + \text{dom } \sigma_{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*} - (\tilde{f} - \lambda c) \right\} = \mathcal{X}. \quad (\text{C.5.45})$$

Take any $g \in \mathcal{X}$ and let us show that it actually belongs to this set. Since $\tilde{f}$ also belongs to $\mathcal{X}$ by assumption, there exists $h \in L_{\heartsuit}^{H_R}$ and some constant $\alpha > 0$ such that, π_0 everywhere

$$g(\xi, \zeta) + \tilde{f}(\xi, \zeta) - h(\xi, \zeta) \leq \alpha(1 + c_0(\xi, \zeta) + c(\xi, \zeta)). \quad (\text{C.5.46})$$

Hence, with $\lambda := \alpha$,

$$g(\xi, \zeta) + \tilde{f}(\xi, \zeta) - h(\xi, \zeta) - \alpha c(\xi, \zeta) \leq \alpha(1 + c(\xi, \xi_0)). \quad (\text{C.5.47})$$

But the RHS belongs to $\text{dom } \sigma_{\mathcal{P}_P(\Xi \times \Xi) \cap L^{H^*}}$ since $\mathbb{E}_{\xi \sim P}[c(\xi, \xi_0)] < +\infty$. Then the LHS is inside the domain, as well. Moreover, h belongs to $L_{\heartsuit}^{H_R}$ so that it lies inside $\text{dom}(R + \lambda S)_* = \frac{1}{1+\lambda\delta/\varepsilon} \text{dom } R$. Hence, we have shown that g belongs to the LHS of (C.5.45), which shows the equality. ■

Lemma C.5.12. *Assume that c is continuous on Ξ^2 . For any $\pi \in \mathcal{P}_P(\Xi \times \Xi)$ such that $\mathbb{E}_{\pi}[c]$ is finite, there exists $\pi_t \in \mathcal{P}_P(\Xi \times \Xi)$ for $t = 1, 2, \dots$ such that*

1. π_t is absolutely continuous w.r.t. π_0 for $t = 1, 2, \dots$,
2. $(\pi_t)_t$ converges weakly to π , i.e., w.r.t. the topology $\sigma(\mathcal{M}(\Xi^2), \mathcal{C}_b(\Xi \times \Xi))$,
3. $\mathbb{E}_{\pi_t}[c] \rightarrow \mathbb{E}_{\pi}[c]$ as $t \rightarrow \infty$.

In particular, for any $g \in L^0(\pi_0)$ l.s.c. such that $g_- \in L^{H_c}$,

$$\mathbb{E}_{\pi}[g] \leq \liminf_{t \rightarrow \infty} \mathbb{E}_{\pi_t}[g], \quad (\text{C.5.48})$$

and, as a consequence, $(\pi)_t$ also converges to π for the topology $\sigma((L_c^H)^*, \mathcal{C}(\Xi^2) \cap L^{H_c})$.

Proof. Step 1: building a partition of Ξ^2 . For convenience, denote by $\|\cdot\|$ a norm on $\mathbb{R}^d$, which contains Ξ and by $\overline{\mathbb{B}}(0, r)$ the closed ball of radius r around 0 in $\mathbb{R}^d$. Fix $\Delta > 0$. By absolute continuity of c on compact sets, for any $n, m \geq 1$, there exists $\eta_{n,m} \in (0, \Delta]$ such that, for any $\xi \in \Xi \cap \overline{\mathbb{B}}(0, n)$, $\zeta, \zeta' \in \Xi \cap \overline{\mathbb{B}}(0, m)$,

$$\|\zeta - \zeta'\| \leq \eta_{n,m} \implies |c(\xi, \zeta') - c(\xi, \zeta)| \leq \Delta. \quad (\text{C.5.49})$$

Fix n and m greater or equal than 1. We consider a finite family of disjoint open sets $\mathcal{U}_i^m$ for $i \in I^m$ which are included in $\mathbb{B}(0, m) \setminus \overline{\mathbb{B}}(0, m-1)$, whose diameters are at most $\eta_{n,m}$ and which satisfy $\overline{\bigcup_{i \in I^m} \mathcal{U}_i^m} \supset \mathbb{B}(0, m) \setminus \overline{\mathbb{B}}(0, m-1)$. Define $J^m = \{i \in I^m : \mathcal{U}_i^m \cap \text{int } \Xi \neq \emptyset\}$. We first show

$$\overline{\bigcup_{i \in J^m} \mathcal{U}_i^m \cap \text{int } \Xi} \supset (\mathbb{B}(0, m) \setminus \overline{\mathbb{B}}(0, m-1)) \cap \Xi \quad (\text{C.5.50})$$

by considering $\zeta \in (\mathbb{B}(0, m) \setminus \overline{\mathbb{B}}(0, m-1)) \cap \Xi$ and $\mathcal{U}$ neighborhood of ζ . We thus have to prove that $\bigcup_{i \in J^m} \mathcal{U}_i^m \cap \text{int } \Xi \cap \mathcal{U}$ is not empty. Since $\mathbb{B}(0, m) \setminus \overline{\mathbb{B}}(0, m-1)$ is an open set that contains ζ , it suffices to show this for some $\mathcal{U} \subset (\mathbb{B}(0, m) \setminus \overline{\mathbb{B}}(0, m-1))$. This means that, in particular, $\mathcal{U} \subset \overline{\bigcup_{i \in I^m} \mathcal{U}_i^m}$. Because ζ belongs to Ξ with $\Xi = \overline{\text{int } \Xi}$, it holds that $\mathcal{U} \cap \text{int } \Xi \neq \emptyset$. Therefore, we have that $\mathcal{U} \cap \text{int } \Xi \cap \overline{\bigcup_{i \in I^m} \mathcal{U}_i^m} \neq \emptyset$. Since $\mathcal{U} \cap \text{int } \Xi$ and $\bigcup_{i \in I^m} \mathcal{U}_i^m$ are both open, this implies that $\mathcal{U} \cap \text{int } \Xi \cap \bigcup_{i \in I^m} \mathcal{U}_i^m \neq \emptyset$, which concludes the proof of (C.5.50). Moreover, since J^m is finite, $\overline{\bigcup_{i \in J^m} \mathcal{U}_i^m \cap \text{int } \Xi} = \bigcup_{i \in J^m} \overline{\mathcal{U}_i^m \cap \text{int } \Xi}$ so that $\bigcup_{i \in J^m} \overline{\mathcal{U}_i^m \cap \text{int } \Xi} \supset (\mathbb{B}(0, m) \setminus \overline{\mathbb{B}}(0, m-1)) \cap \Xi$ holds as well.

Still keeping $n \geq 1$ fixed, we have built an (at most) countable number of disjoint open sets $(\mathcal{V}_i)_{i \leq M_n} := (\mathcal{U}_i^m \cap \text{int } \Xi)_{i \in J^m, m \geq 1}$, with $M_n \in \mathbb{N} \cup \{+\infty\}$ such that for any $i \leq M_n$, there is some $m \geq 1$ such that $\mathcal{V}_i \subset \mathbb{B}(0, m) \setminus \overline{\mathbb{B}}(0, m-1) \cap \text{int } \Xi$ and such that the diameter of $\mathcal{V}_i$ is at most $\eta_{n,m}$. Moreover, $\bigcup_{i \in I} \overline{\mathcal{V}_i}$ is the whole of Ξ . We build a partition of Ξ by recursively defining

$$\mathcal{W}_i^n = \overline{\mathcal{V}_i} \setminus \bigcup_{j \leq i-1} \mathcal{W}_j^n = \mathcal{V}_i \cup \left(\partial \mathcal{V}_i \setminus \bigcup_{j \leq i-1} \mathcal{W}_j^n \right) \text{ for } i \leq M_n. \quad (\text{C.5.51})$$

Therefore, we have built an (at most) countable number of disjoint measurable sets $\mathcal{W}_i^n$ for $i \leq M_n$, whose union is Ξ and such that for any $i \leq M_n$, there is some $m \geq 1$ such that $\mathcal{W}_i^n \subset \overline{\mathbb{B}}(0, m) \cap \Xi$. Moreover, the diameter of each $\mathcal{W}_i^n$ is at most $\eta_{n,m}$ and each of these sets has non-empty interior.

The full partition of Ξ^2 that we finally consider is made of the sets $((\mathbb{B}(0, n) \setminus \mathbb{B}(0, n-1)) \cap \Xi) \times \mathcal{W}_i^n$ for $n \geq 1$ and $i \leq M_n$. To summarize, we have built, for $\Delta > 0$, a partition of the form $(A_i^\Delta \times B_j^\Delta)_{i \in I, j \in J(i)}$ which is a (at most) countable and measurable partition of Ξ^2 satisfying:

1. $(A_i^\Delta)_{i \in I}$ is a partition of Ξ and, for any $i \in I$, $(B_j^\Delta)_{j \in J(i)}$ is a partition of Ξ .
2. For $i \in I$, $j \in J(i)$, $\xi \in A_i^\Delta$, and $\zeta, \zeta' \in B_j^\Delta$, we have $\|\zeta - \zeta'\| \leq \Delta$ and $|c(\xi, \zeta') - c(\xi, \zeta)| \leq \Delta$.
3. For any $i \in I$, $j \in J(i)$, B_j^Δ has non-empty interior.

Step 2: construction of the sequence of measures. For $\Delta > 0$, we define the measure $\pi^\Delta \in \mathcal{P}_p(\Xi \times \Xi)$ for any $\xi \in \Xi$ and any measurable $B \subset \Xi$ by

$$\pi^\Delta(B|\xi) := \sum_{i \in I} \mathbb{1}_{\xi \in A_i^\Delta} \sum_{j \in J(i)} \frac{\pi(B_j^\Delta|\xi)}{\pi_0(B_j^\Delta|\xi)} \pi_0(B \cap B_j^\Delta|\xi), \quad (\text{C.5.52})$$

which is well-defined because of the B_j^Δ having non-empty interiors. Moreover, by construction, it is absolutely continuous w.r.t. π_0 , so that π^Δ satisfies (1).

For $i \in I$ and $j \in J(i)$, choose $\zeta_{i,j}^\Delta \in B_j^\Delta$. By the Portmanteau theorem, see [Klenke \(2014, Th 13.16\)](#), it suffices to show that $\int g \, d\pi^\Delta$ converges to $\int g \, d\pi$ as $\Delta \rightarrow 0$ for any bounded Lipschitz function g to prove the weak convergence of the sequence of measures (π^Δ) to π . Thus, take a bounded 1-Lipschitz function $g : \Xi^2 \rightarrow \mathbb{R}$, then

$$\left| \int g \, d\pi^\Delta - \int g \, d\pi \right| \leq \mathbb{E}_{\xi \sim \mathbb{P}} \left[\sum_{i \in I} \mathbb{1}_{\xi \in A_i^\Delta} \sum_{j \in J(i)} \left| \frac{\pi(B_j^\Delta|\xi)}{\pi_0(B_j^\Delta|\xi)} \int_{B_j^\Delta} g(\xi, \zeta) \pi_0(d\zeta|\xi) - \int_{B_j^\Delta} g(\xi, \zeta) \pi(d\zeta|\xi) \right| \right] \quad (\text{C.5.53})$$

$$\leq \mathbb{E}_{\xi \sim \mathbb{P}} \left[\sum_{i \in I} \mathbb{1}_{\xi \in A_i^\Delta} \sum_{j \in J(i)} \pi(B_j^\Delta|\xi) \left(\frac{1}{\pi_0(B_j^\Delta|\xi)} \int_{B_j^\Delta} |g(\xi, \zeta) - g(\xi, \zeta_{i,j}^\Delta)| \pi_0(d\zeta|\xi) \right) \right] \quad (\text{C.5.54})$$

$$+ \mathbb{E}_{\xi \sim \mathbb{P}} \left[\sum_{i \in I} \mathbb{1}_{\xi \in A_i^\Delta} \sum_{j \in J(i)} \pi(B_j^\Delta|\xi) \left(\frac{1}{\pi(B_j^\Delta|\xi)} \int_{B_j^\Delta} |g(\xi, \zeta) - g(\xi, \zeta_{i,j}^\Delta)| \pi(d\zeta|\xi) \right) \right] \quad (\text{C.5.55})$$

$$\leq \mathbb{E}_{\xi \sim P} \left[\sum_{i \in I} \mathbb{1}_{\xi \in A_i^\Delta} \sum_{j \in J(i)} \pi(B_j^\Delta | \xi) \left(\frac{1}{\pi_0(B_j^\Delta | \xi)} \int_{B_j^\Delta} \|\zeta - \zeta_{i,j}^\Delta\| \pi_0(d\zeta | \xi) \right) \right] \quad (\text{C.5.56})$$

$$+ \mathbb{E}_{\xi \sim P} \left[\sum_{i \in I} \mathbb{1}_{\xi \in A_i^\Delta} \sum_{j \in J(i)} \pi(B_j^\Delta | \xi) \left(\frac{1}{\pi(B_j^\Delta | \xi)} \int_{B_j^\Delta} \|\zeta - \zeta_{i,j}^\Delta\| \pi(d\zeta | \xi) \right) \right] \quad (\text{C.5.57})$$

$$\leq 2 \mathbb{E}_{\xi \sim P} \left[\sum_{i \in I} \mathbb{1}_{\xi \in A_i^\Delta} \sum_{j \in J(i)} \pi(B_j^\Delta | \xi) \Delta \right] = 2\Delta, \quad (\text{C.5.58})$$

which vanishes as Δ goes to 0. This proves point (2) of the result.

Using the same decomposition of the integral, we get the same inequality for c ,

$$\left| \int c \, d\pi^\Delta - \int c \, d\pi \right| \leq \Delta, \quad (\text{C.5.59})$$

which shows the convergence of the integral $\mathbb{E}_{\pi^\Delta}[c]$ to $\mathbb{E}_\pi[c]$, which is point (3) of the result.

Finally, take $g \in L^0(\pi_0)$ l.s.c. and such that $g_- \in L^{H_c}$, this means that there exists $\alpha > 0$ such that, π_0 a.s., $g \geq -\alpha(1 + c_0 + c)$. But by construction, $\mathbb{E}_{\pi^\Delta}[1 + c_0 + c] = 1 + \mathbb{E}_P[c(\cdot, \xi_0)] + \mathbb{E}_{\pi^\Delta}[c]$ which goes to $\mathbb{E}_\pi[1 + c_0 + c]$ as $\Delta \rightarrow 0$. The last part of the result then follows from Villani (2008, Lemma 4.3) with $c \leftarrow -\alpha(1 + c_0 + c)$ and $h \leftarrow g$. ■

Lemma C.5.13. *Under the assumptions of Theorem C.6, the values of the two problems coincide:*

$$\sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \overline{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}, \langle \pi \rangle c + S(\pi) \leq \rho \right\} \quad (\text{C.5.60})$$

and

$$\sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi), \langle \pi \rangle c + S(\pi) \leq \rho \right\}. \quad (\text{C.5.61})$$

Proof. We divide the proof in two parts, based on the two regularity assumption in Theorem C.6.

Case (a): R is $\sigma(\mathcal{X}^*, \mathcal{C}(\Xi^2) \cap L^{H_c})$ -u.s.c., f is l.s.c. and $\tilde{f}_- \in L^{H_c}$. Thanks to the existence of a strictly feasible point π_S ($\langle \pi_S \rangle c + S(\pi_S) < \rho$) and the convexity (resp. concavity) of the constraint (resp. objective), Lemma C.5.9 ensures that the values of the two problems do not change when the inequality constraints are replaced by strict inequality constraints.

The value of (C.5.60) is thus equal to

$$\sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \overline{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*}, \langle \pi \rangle c + S(\pi) < \rho \right\} \quad (\text{C.5.62})$$

$$= \sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) - \iota_{\langle \cdot \rangle c + S < \rho}(\pi) : \pi \in \overline{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*} \right\}, \quad (\text{C.5.63})$$

with $\langle \cdot \rangle \tilde{f} - R - \iota_{\langle \cdot \rangle c + S < \rho}$ now $\sigma(\mathcal{X}^*, \mathcal{X})$ -l.s.c. since $\mathcal{C}(\Xi^2) \cap L^{H_c} \subset \mathcal{X}$ so that $\sigma(\mathcal{X}^*, \mathcal{C}(\Xi^2) \cap L^{H_c})$ is weaker than $\sigma(\mathcal{X}^*, \mathcal{X})$. Thus, the weak-* closure can be removed, and (C.5.60) is equal to

$$\sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) - \iota_{\langle \cdot \rangle c + S < \rho}(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^* \right\} \quad (\text{C.5.64})$$

$$= \sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*, \langle \pi \rangle c + S(\pi) < \rho \right\} \quad (\text{C.5.65})$$

$$= \sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0), \langle \pi \rangle c + S(\pi) < \rho \right\}, \quad (\text{C.5.66})$$

where the last equality follows from [Lemma C.5.5](#) since $S(\pi)$ and $\langle \pi \rangle c$ have to be both finite for π to be feasible and $R \propto S$. All that is left to show is the equality

$$\sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0), \langle \pi \rangle c + S(\pi) < \rho \right\} \quad (\text{C.5.67})$$

$$= \sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi), \langle \pi \rangle c + S(\pi) < \rho \right\}. \quad (\text{C.5.68})$$

The inequality ($\leq$) is straightforward. Concerning the other side, for any $\pi \in \mathcal{P}_P(\Xi \times \Xi)$ such that $\langle \pi \rangle c + S(\pi) < \rho$, we can consider $(\pi_t)_t$ the sequence of probability distributions in $\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0)$ given by [Lemma C.5.12](#). Thanks to c belonging to $\mathcal{C}(\Xi^2) \cap L^{H_c}$ and the u.s.c. assumption on $S \propto R$, $\limsup_{t \rightarrow \infty} \langle \pi_t \rangle c + S(\pi_t) \leq \langle \pi \rangle c + S(\pi) < \rho$ showing that π_t is strictly feasible after some time. Moreover, since $\tilde{f}$ is l.s.c. and $\tilde{f}_-$ belongs to L^{H_c} , we get that

$$\langle \pi \rangle \tilde{f} - R(\pi) \leq \liminf_{t \rightarrow \infty} \langle \pi_t \rangle \tilde{f} - R(\pi_t) \quad (\text{C.5.69})$$

$$\leq \sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0), \langle \pi \rangle c + S(\pi) < \rho \right\}, \quad (\text{C.5.70})$$

which achieves the proof.

Case (2): $\text{dom } R \subset L^0(\pi_0)$ and, for any sequence $g_t \in \text{dom } H_R$ non-negative such that $g_t \rightarrow 0$ as $t \rightarrow \infty$, $H_R(g_t) \rightarrow 0$. In this case, H_R satisfies (H4) of [Pennanen and Perkkiö \(2022, Thm. 22\)](#) so that $(L_\infty^{H_R})^* = L^{H_R^*}$. As a consequence, $\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*$ is $\sigma(\mathcal{X}^*, \mathcal{X})$ closed so that

$$\overline{\mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*} = \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0) \cap \mathcal{X}^*.$$

Therefore, using [Lemma C.5.5](#), one gets that

$$\sup \left\{ \langle \pi \rangle \tilde{f} - R(\pi) : \pi \in \mathcal{P}_P(\Xi \times \Xi) \cap L^0(\pi_0), \langle \pi \rangle c + S(\pi) \leq \rho \right\}, \quad (\text{C.5.71})$$

and the result follows from the assumption on the domain of R . $\blacksquare$

C.5.5 Examples

We finally develop the two examples of [Appendix C.2.2](#) by choosing carefully the coupling π_0 .

Corollary C.5.14 (Cost-regularized WDRO). *Let [Assumption C.3](#) holds. Suppose that the function $f : \Xi \rightarrow \mathbb{R}_+$ in $L^1(P)$, l.s.c., and such that the function $(\xi, \zeta) \mapsto$*

$f(\zeta)/(1 + c_0(\xi, \zeta) + c(\xi, \zeta))$ is bounded, above and below, over Ξ^2 . Then, for any $\varepsilon, \delta > 0$, the problem

$$\sup \{ \mathbb{E}_{\pi_2}[f] - \varepsilon \mathbb{E}_{\pi}[c] : \pi \in \mathcal{P}_P(\Xi \times \Xi), \mathbb{E}_{\pi}[c] + \delta \mathbb{E}_{\pi}[c] \leq \rho \} \quad (\text{C.5.72})$$

coincides with its dual

$$\inf \left\{ \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\operatorname{ess\,sup}_{Leb} \{ f - (\varepsilon + (1 + \delta)\lambda)c(\xi, \cdot) \} \right] : \lambda \geq 0 \right\}. \quad (\text{C.5.73})$$

Proof. In order to prove this result, we choose $\pi_0 \in \mathcal{P}_P(\Xi \times \Xi)$ such that, for any $\xi \in \Xi$,

$$\pi_0(d\zeta|\xi) \propto \frac{\mathbb{1}_{\Xi}(\zeta)}{1 + c_0(\xi, \zeta) + c(\xi, \zeta)} e^{-\|\zeta\|^2/2} d\zeta. \quad (\text{C.5.74})$$

We also define R for all $\pi \in (L^b(\Xi^2))^*$,

$$R(\pi) = \begin{cases} \varepsilon \langle \pi \rangle c & \text{if } \pi \in (L_c^H)^* \text{ non-negative} \\ \varepsilon \int c \, d\pi & \text{if } \pi \in \mathcal{M}(\Xi^2) \text{ non-negative} \\ +\infty & \text{otherwise.} \end{cases} \quad (\text{C.5.75})$$

With this definition, R satisfies the assumption of [Lemma C.5.3](#) with $\pi_R = 0$ and we have

$$H_R(g) = \sup_{\pi \in L^0(\pi_0): \pi \geq 0} \int (|g| - \varepsilon c) \pi \, d\pi_0 = \iota_{|g| \leq \varepsilon c}, \quad (\text{C.5.76})$$

and thus $L_{\diamond}^{H_R} = \{0\}$ so that $\mathcal{X}$ is equal to L^{H_c} .

It remains to check the regularity conditions on R . $R|_{\mathcal{X}^*}$ is $\sigma(\mathcal{X}^*, \mathcal{X})$ -l.s.c. since c belongs to $\mathcal{X}$ and the set of non-negative linear forms in $\mathcal{X}^*$ is $\sigma(\mathcal{X}^*, \mathcal{X})$ -closed. On $\mathcal{P}_P(\Xi \times \Xi)$, R coincides with $\pi \mapsto \varepsilon \int c \, d\pi$ and c is inside $\mathcal{C}(\Xi^2) \cap L^{H_c}$ so it is actually $\sigma(\mathcal{X}^*, \mathcal{C}(\Xi^2) \cap L^{H_c})$ continuous and thus regularity condition (a) of [Theorem C.6](#) holds. Finally, the strict feasibility condition is satisfied thanks to the distribution $\pi(d\xi, d\zeta) = P(d\xi)\delta_{\xi}(d\zeta)$ so [Theorem C.6](#) applies. Since $L^{\infty}(\pi_0) \subset \mathcal{X}^*$, the preconjugates of $R + \lambda S$ is given by,

$$(R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*})_*(g) = \iota_{g \leq (\varepsilon + \lambda \delta)c}. \quad (\text{C.5.77})$$

Hence, since the function $(\varepsilon + \lambda \delta)c$ belongs to $\mathcal{X} = L^{H_c}$,

$$\inf \left\{ \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\operatorname{ess\,sup}_{\pi_0(\cdot|\xi)} \{ f - \lambda c(\xi, \cdot) - g(\xi, \cdot) \} \right] + (R|_{\mathcal{X}^*} + \lambda S|_{\mathcal{X}^*})_*(g) : g \in \mathcal{X} \right\} \quad (\text{C.5.78})$$

$$= \mathbb{E}_{\xi \sim P} \left[\operatorname{ess\,sup}_{Leb} \{ f - (\varepsilon + (1 + \delta)\lambda)c(\xi, \cdot) \} \right], \quad (\text{C.5.79})$$

which gives the result. $\blacksquare$

Corollary C.5.15 (ϕ -divergence-regularized WDRO). *Let [Assumption C.3](#) hold and take $\varepsilon, \delta > 0$. Consider $\pi_0 \in \mathcal{P}(\Xi \times \Xi)$ with $\mathbb{E}_{\pi_0}[c] < +\infty$, a convex l.s.c.*

function $\phi : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ with such that $\phi(1) = 0$, $[1, +\infty) \subset \text{dom } \phi \subset [0, +\infty)$, $\phi'(+\infty) = +\infty$ and define the associated divergence

$$\forall \pi \in \mathcal{P}(\Xi \times \Xi), \quad D_\phi(\pi|\pi_0) := \begin{cases} \int_{\Xi^2} \phi\left(\frac{d\pi}{d\pi_0}\right) d\pi_0 & \text{if } \pi \text{ is absolutely continuous w.r.t. } \pi_0 \\ +\infty & \text{otherwise.} \end{cases} \quad (\text{C.5.80})$$

Assume also that f satisfies the growth condition:

$$\exists g \in L^{H_c}, \forall \alpha > 0, \int_{\Xi^2} \phi^*(\alpha|\tilde{f} - g|) d\pi_0 < +\infty. \quad (\text{C.5.81})$$

Then, with $R \leftarrow \varepsilon D_\phi(\cdot|\pi_0)$ and $S \leftarrow \delta D_\phi(\cdot|\pi_0)$, if (R-WDRO) is strictly feasible, its value is equal to

$$\inf_{\lambda \geq 0} \inf_{\psi \in \mathcal{C}(\Xi^2)} \lambda \rho + \mathbb{E}_{\xi \sim P} \left[\sup_{\zeta \in \Xi} f(\zeta) - \lambda c(\xi, \zeta) - \psi(\xi, \zeta) \right] + (\varepsilon + \lambda \delta) \int_{\Xi^2} \phi^*\left(\frac{\psi(\xi, \zeta)}{\varepsilon + \lambda \delta}\right) d\pi_0(\xi, \zeta). \quad (\text{C.5.82})$$

Proof. D_ϕ admits the following variational formula, see e.g., [Agrawal and Horel \(2021, Prop. 4.2.8\)](#) applied with the pair $(L^\infty(\pi_0), \mathcal{M}(\Xi^2))$,

$$\forall \pi \in \mathcal{M}(\Xi^2), \quad D_\phi(\pi|\pi_0) = \sup \left\{ \langle \pi, \psi \rangle - \int_{\Xi^2} \phi^* \circ \psi d\pi_0 : \psi \in L^\infty(\pi_0) \right\}. \quad (\text{C.5.83})$$

As seen in the second case of [Lemma C.5.13](#), $\text{dom } R \subset L^0(\pi_0)$ so that $\mathcal{X}^* \subset L^0(\pi_0)$. Hence, this variational formula applies to the whole $\mathcal{X}^*$ and since $L^\infty(\pi_0) \subset L^{H_c} \subset \mathcal{X}$, $D_\phi(\cdot|\pi_0)$ is $\sigma(\mathcal{X}^*, \mathcal{X})$ -l.s.c.. Moreover, [Agrawal and Horel \(2021, Prop. 4.2.6\)](#) with the $(\mathcal{X}, \mathcal{M}(\Xi^2))$ and $(\mathcal{X}, \mathcal{X}^*)$ gives

$$H_R(g) = (R|_{\mathcal{X}^*})_*(|g|) = \int_{\Xi^2} \phi^* \circ g d\pi_0. \quad (\text{C.5.84})$$

The continuity at 0 of H_R follows from the dominated convergence theorem and the fact that ϕ^* is non-decreasing. The growth condition means that $\tilde{f}$ lies in $\mathcal{X}$, and we can apply [Theorem C.6](#). $\blacksquare$

EXACT GENERALIZATION GUARANTEES FOR
(REGULARIZED) WASSERSTEIN DISTRIBUTIONALLY
ROBUST MODELS

This appendix contains the proofs for [Chapter 6](#). It is extracted from [Azizian et al. \(2023c\)](#).

We provide here the proofs of our main results [Theorems 6.1, 6.2 and 6.6](#), along with detailed versions that include explicit bounds. We start, in [Appendix D.1](#), by referencing preliminary results, reformulate some of our assumptions, and introduce quantities that appear in the final bounds.

As sketched in [§ 6.3.2](#), our proof is built on two main parts. The first part is to obtain a concentration bound on the dual problems [\(6.19\)](#) and [\(6.20\)](#) by leveraging on a lower bound on the dual multiplier. This concentration result is presented in [Appendix D.2](#) where we assume that such a lower-bound is given. The second part of the proof then consists in establishing the lower-bound. We have to distinguish two cases: when ρ is small ([Appendix D.3](#)) and when ρ is close to the critical radius ρ_c ([Appendix D.4](#)). For the latter, we also need to treat separately the cases where the WDRO problem is regularized or not (respectively [Appendix D.4.2](#) and [Appendix D.4.1](#)): this is where the [Assumptions 6.4 and 6.5](#), that are not required in the regularized case, come into play. Putting together these two parts, we obtain our precise theorems in [Appendix D.5](#) and show how they imply our main results. [Appendix D.6](#) then complements our theorems to obtain [Theorem 6.3](#). Finally, some variations of known results and technical computations are compiled in [Appendix D.7](#) as standalone lemmas and, in the numerical illustrations below, we provide numerical illustrations for linear models.

D.1 PRELIMINARIES

This section presents preliminary results before we start the proofs in [Appendix D.2](#). In the first part of this section [Appendix D.1.1](#), we present a weaker and more detailed version of [Assumption 6.2](#), namely [Assumption D.1](#), that will suffice for all the proofs in the appendix. We also introduce several quantities that will appear in the final bounds. Then, in [Appendix D.1.1](#) we recall the dual problems introduced in [§ 6.3.2](#) and justify that strong duality holds. Preliminary approximation results on the dual are then given in [Appendix D.1.3](#). We then proceed to show the relative compactness of the class $\mathcal{F}$ w.r.t. several metrics in [Appendix D.1.4](#). These properties provide a convenient way of ensuring that quantities involving $\mathcal{F}$ are finite, e.g., complexity measures or supremums over $\mathcal{F}$. Finally, we introduce the so-called parametric Morse-Bott condition of [Arbel and Mairal \(2022\)](#) and show how it implies the second item of [Assumption 6.2](#) in a Riemannian setting.

D.1.1 Detailed assumption on the function class and important quantities

Here we present the precise assumptions that we will refer to in the proofs. While [Assumptions 6.1](#) and [6.3](#) are used as presented in the main text, we slightly weaken [Assumption 6.2](#) to [Assumption D.1](#). We also introduce some quantities that we will be of interest for the proofs and the final results.

Assumption D.1 (On the function class). Consider $\mathcal{F}$ a set of real-valued non-negative continuous functions on Ξ . We assume that:

- the functions $f \in \mathcal{F}$ are uniformly L_2 -smooth;
- the gradients are uniformly bounded, i.e.,

$$G := \sup_{f \in \mathcal{F}} \sup_{\xi \in \text{supp } P} \|\nabla f(\xi)\|_2 < +\infty \quad (\text{D.1.1})$$

- when $\varepsilon = 0$, the supremum in [\(6.18\)](#) is finite, i.e.,

$$\tilde{F}(\lambda) := \sup_{f \in \mathcal{F}} \sup_{\xi \in \text{supp } P} \sup_{\arg \max \{f - \frac{\lambda}{2} \|\xi - \cdot\|^2\}} f < +\infty. \quad (\text{D.1.2})$$

Note that the non-negativity assumption is without loss of generality since otherwise, it suffices to consider $\tilde{\mathcal{F}} := \{f - \min f : f \in \mathcal{F}\}$ and our results are invariant by addition of a constant.

The blanket assumptions for the remaining of the appendix will be [Assumptions 6.1, 6.3](#) and [D.1](#).

The following finite quantities are relevant for the proofs and appear in the quantitative versions of [Theorems 6.1, 6.2](#) and [6.6](#).

$$C^* := \sup \left\{ \frac{1}{2} \|\xi - \xi^*\|^2 : \xi \in \text{supp } P, f \in \mathcal{F}, \xi^* \in \arg \max f \right\}, \quad (\text{D.1.3})$$

$$C(\sigma) := \sup_{\xi \in \text{supp } P} \mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right], \quad (\text{D.1.4})$$

$$C_{\mathcal{F}}(\varepsilon, \sigma) := \sup_{f \in \text{conv}(\mathcal{F})} \sup_{\xi \in \text{supp } P} \mathbb{E}_{\zeta \sim \pi_\sigma^{f/\varepsilon}(\cdot|\xi)} \frac{1}{2} \|\xi - \zeta\|^2, \quad (\text{D.1.5})$$

$$\text{and } \text{Var}(\varepsilon, \sigma) := \sup_{f \in \text{conv}(\mathcal{F})} \sup_{\xi \in \text{supp } P} \sup_{\lambda \geq 0} \text{Var}_{\zeta \sim \pi_\sigma^{\frac{f - \lambda \|\xi - \cdot\|^2/2}{\varepsilon}}(\cdot|\xi)} \frac{1}{2} \|\xi - \zeta\|^2, \quad (\text{D.1.6})$$

where $\varepsilon > 0$, $\sigma > 0$, and π_σ is given by [\(6.30\)](#).

D.1.2 Strong duality

As mentioned in [§ 6.3.2](#), duality plays a central role in our proofs. Let us recall the central notion of dual generator functions, introduced in [\(6.18\)](#): for any $f \in \mathcal{F}$, $\xi \in \Xi$, $\lambda, \varepsilon \geq 0$ and $\sigma > 0$, the dual generator ϕ is given as

$$\phi(f, \xi, \lambda, \varepsilon, \sigma) := \begin{cases} \sup_{\zeta \in \Xi} \{f(\zeta) - \frac{\lambda}{2} \|\xi - \zeta\|^2\} & \text{if } \varepsilon = 0 \\ \varepsilon \log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \exp \left(\frac{f(\zeta) - \lambda \|\xi - \zeta\|^2/2}{\varepsilon} \right) \right) & \text{if } \varepsilon > 0. \end{cases} \quad (\text{D.1.7})$$

Our proofs are based on the (strong) dual formulations of WDRO, as given by the following lemma that summarizes results of the literature for the regularized and unregularized cases.

Lemma D.1.1. *Under the blanket assumptions, for $f \in \mathcal{F}$, $\rho > 0$, $\varepsilon \geq 0$ and $\sigma > 0$,*

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\lambda \geq 0} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbf{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.1.8})$$

$$\text{and } \mathcal{R}_{\rho^2}^\varepsilon(f) = \inf_{\lambda \geq 0} \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbf{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] . \quad (\text{D.1.9})$$

Proof. See [Blanchet and Murthy \(2019\)](#), [Gao and Kleywegt \(2022\)](#) for the unregularized case and [Azizian et al. \(2023a\)](#) for the regularized case. ■

D.1.3 Approximation of the dual generator ϕ

Important preliminary results for our upcoming concentration bounds (in [Appendix D.2](#)) are quantitative approximations of the dual generator ϕ , namely [Proposition D.1.2](#) and [Lemma D.1.3](#). In particular, these results also imply bounds on ϕ in [Corollary D.1.5](#).

Proposition D.1.2 (Bounding the distance between ϕ and f). *There are positive constants $\lambda_1, \varepsilon_1, \sigma_1, c_1, c_2$ which depend on G, R and d such that taking some $\bar{\lambda} \geq \underline{\lambda} \geq \lambda_1 + L_2$, we have for any $f \in \mathcal{F}$, $\xi \in \text{supp } \mathbf{P}$, $\lambda \in [\underline{\lambda}, \bar{\lambda}]$, $\varepsilon \in [0, \varepsilon_1]$ and $\sigma \in (0, \sigma_1]$*

$$|\phi(f, \xi, \lambda, \varepsilon, \sigma) - f(\xi)| \leq M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \quad (\text{D.1.10})$$

where

$$M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) := \frac{1}{2\underline{\lambda}} G^2 + \frac{\varepsilon d}{2} \log \left(\frac{\bar{\lambda}}{\varepsilon} + \frac{1}{\sigma^2} \right) + \varepsilon \log 2 + \varepsilon d |\log \sigma| + \varepsilon c_1 e^{-c_2 \left(\frac{\bar{\lambda} - L_2}{\varepsilon} \right)^{\frac{1}{3}}} . \quad (\text{D.1.11})$$

The proof of this result is based on the following second approximation result which gives a precise approximation of ϕ that will be used several times in the upcoming proof.

More precisely, we want to approximate ϕ by a Taylor development $\bar{\phi}$ defined for any $f \in \mathcal{F}$, $\xi \in \text{supp } \mathbf{P}$, $\lambda \geq 0$, $\varepsilon \geq 0$ and $\sigma > 0$ as

$$\bar{\phi}(f, \xi, \lambda, \varepsilon, \sigma) := f(\xi) + \frac{1}{2 \left(\lambda + \frac{\varepsilon}{\sigma^2} \right)} \|\nabla f(\xi)\|_2^2 - \frac{\varepsilon d}{2} \log \left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2} \right) + \varepsilon \log \frac{(2\pi)^{\frac{d}{2}}}{Z(\xi, \sigma)} \quad (\text{D.1.12})$$

where $Z(\xi, \sigma) := \int_{\Xi} e^{-\frac{\|\xi - \zeta\|_2^2}{2\sigma^2}} d\zeta$. The distance between ϕ and $\bar{\phi}$ is then controlled by the following Laplace approximation lemma.

Lemma D.1.3 (Approximation of ϕ). *There are positive constants $\lambda_1, \varepsilon_1, \sigma_1, c_1, c_2$ which depend on G, R and d such that $\varepsilon_1 \leq \lambda_1$ and, when $\varepsilon \in [0, \varepsilon_1]$, $\sigma \in (0, \sigma_1]$ and $\lambda \geq \lambda_1 + L_2$, we have for any $f \in \mathcal{F}$, $\xi \in \text{supp } \mathbf{P}$*

$$\bar{\phi}(f, \xi, \lambda + L_2, \varepsilon, \sigma) - \varepsilon c_1 e^{-c_2 \left(\frac{\lambda + L_2}{\varepsilon} \right)^{\frac{1}{3}}} \leq \phi(f, \xi, \lambda, \varepsilon, \sigma) \leq \bar{\phi}(f, \xi, \lambda - L_2, \varepsilon, \sigma) + \varepsilon c_1 e^{-c_2 \left(\frac{\lambda - L_2}{\varepsilon} \right)^{\frac{1}{3}}} . \quad (\text{D.1.13})$$

Proof. Fix $f \in \mathcal{F}$, $\xi \in \text{supp } P$, $\lambda \geq 0$, $\varepsilon \geq 0$ and $\sigma > 0$. To bound the error between ϕ and its approximation $\bar{\phi}$, we introduce an intermediate approximation $\tilde{\phi}$ defined as

$$\tilde{\phi}(f, \xi, \lambda, \varepsilon, \sigma) := \begin{cases} \varepsilon \log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \exp \left(\frac{f(\xi) + \langle \nabla f(\xi), \zeta - \xi \rangle - \lambda \|\xi - \zeta\|^2/2}{\varepsilon} \right) \right) & \text{if } \varepsilon > 0 \\ \sup_{\zeta \in \Xi} [f(\xi) + \langle \nabla f(\xi), \zeta - \xi \rangle - \frac{\lambda}{2} \|\xi - \zeta\|^2] & \text{if } \varepsilon = 0, \end{cases} \quad (\text{D.1.14})$$

which corresponds to ϕ applied to the Taylor approximation of f at ξ (instead of f itself). By smoothness of the functions in $\mathcal{F}$ ([Assumption D.1](#)), we readily have that,

$$\tilde{\phi}(f, \xi, \lambda + L_2, \varepsilon, \sigma) \leq \phi(f, \xi, \lambda, \varepsilon, \sigma) \leq \tilde{\phi}(f, \xi, \lambda - L_2, \varepsilon, \sigma). \quad (\text{D.1.15})$$

Now, all that is left to bound, is the error between $\tilde{\phi}$ and $\bar{\phi}$. Consider first the case where $\varepsilon > 0$ and let us rewrite $\tilde{\phi}$ by using the definition of π_σ :

$$\tilde{\phi}(f, \xi, \lambda, \varepsilon, \sigma) = \varepsilon \log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \exp \left(\frac{f(\xi) + \langle \nabla f(\xi), \zeta - \xi \rangle - \lambda \|\xi - \zeta\|^2/2}{\varepsilon} \right) \right) \quad (\text{D.1.16})$$

$$= \varepsilon \log \left(\int_{\Xi} \exp \left(\frac{1}{\varepsilon} \left(f(\xi) + \langle \nabla f(\xi), \zeta - \xi \rangle - \left(\lambda + \frac{\varepsilon}{\sigma^2} \right) \frac{1}{2} \|\xi - \zeta\|^2 \right) \right) d\zeta \right) - \varepsilon \log Z(\xi, \sigma). \quad (\text{D.1.17})$$

But, looking at the inner expression, we have that

$$f(\xi) + \langle \nabla f(\xi), \zeta - \xi \rangle - \left(\lambda + \frac{\sigma^2}{\varepsilon} \right) \frac{1}{2} \|\xi - \zeta\|^2 = f(\xi) + \frac{1}{2(\lambda + \frac{\sigma^2}{\varepsilon})} \|\nabla f(\xi)\|_2^2 - \frac{1}{2\tau} \|\zeta - \zeta^*(\tau)\|_2^2, \quad (\text{D.1.18})$$

where we defined $\frac{1}{\tau} := \lambda + \frac{\varepsilon}{\sigma^2}$ and $\zeta^*(\tau) = \xi + \tau \nabla f(\xi)$. Hence,

$$\tilde{\phi}(f, \xi, \lambda, \varepsilon, \sigma) = \bar{\phi}(f, \xi, \lambda, \varepsilon, \sigma) + \varepsilon \log \left(\int_{\Xi} \exp \left(-\frac{1}{2\varepsilon\tau} \|\zeta - \zeta^*(\tau)\|_2^2 \right) d\zeta \right) - \frac{\varepsilon d}{2} \log(2\pi\varepsilon\tau). \quad (\text{D.1.19})$$

Define $\lambda'_1 := \frac{\sqrt{6}G}{R}$ and $\tau_1 := \frac{1}{\lambda'_1} = \frac{R}{2G}$ so that $\lambda \geq \lambda'_1$ implies that $\tau \leq \frac{1}{\lambda'_1} = \tau_1$. Let us now check that the conditions of [Lemma D.7.1](#) are satisfied.

1. Since $\zeta^*(0) = \xi$, by [Assumption 6.3](#), $\bar{\mathbb{B}}(\zeta^*(0), R)$ is contained in Ξ .

2. For $\tau \leq \tau_1$, we have that $\tau^2 \|\nabla f(\xi)\|_2^2 \leq \frac{G^2}{(\lambda'_1)^2} = \frac{R^2}{6}$ by definition.

Hence, we can apply [Lemma D.7.1](#) to get that, for any $\lambda \geq \lambda'_1$

$$\varepsilon \log \left(1 - 6^{d/2} e^{-\frac{R^2}{12\varepsilon\tau}} \right) \leq \tilde{\phi}(f, \xi, \lambda, \varepsilon, \sigma) - \bar{\phi}(f, \xi, \lambda, \varepsilon, \sigma) \leq \varepsilon \log \left(1 + 6^{d/2} e^{-\frac{R^2}{12\varepsilon\tau}} \right). \quad (\text{D.1.20})$$

Now, using [Lemma D.7.5](#), we get that there are positive constants $\varepsilon_1, \lambda_1, c_1, c_2$ depending on R, G and d such that, if $\varepsilon \leq \varepsilon_1$ and $\lambda \geq \lambda_1$, then

$$|\tilde{\phi}(f, \xi, \lambda, \varepsilon, \sigma) - \bar{\phi}(f, \xi, \lambda, \varepsilon, \sigma)| \leq \varepsilon c_1 e^{-c_2 \left(\frac{\lambda}{\varepsilon} \right)^{\frac{1}{3}}}. \quad (\text{D.1.21})$$

Moreover, ε_1 can be reduced so that it is less than λ_1 if it is not the case originally.

To finish the proof, let us now come back to the case $\varepsilon = 0$. First, note that (D.1.18) is still valid even with $\varepsilon = 0$ so that we have

$$\tilde{\phi}(f, \xi, \lambda, 0) = \bar{\phi}(f, \xi, \lambda, 0) - \frac{1}{2\tau} \inf_{\zeta \in \Xi} \|\zeta - \zeta^*(\tau)\|_2^2. \quad (\text{D.1.22})$$

But as seen above, for $\tau = \lambda^{-1} \leq \tau_1$, $\zeta^*(\tau)$ is inside $\bar{\mathbb{B}}(\zeta^*(0), R)$ so that $\tilde{\phi}(f, \xi, \lambda, 0) = \bar{\phi}(f, \xi, \lambda, 0)$.

We conclude the proof by noticing that the obtained bounds are valid for any $0 < \varepsilon \leq \varepsilon_1$, $\lambda \geq \lambda_1 + L_2$, $f \in \mathcal{F}$ and $\xi \in \text{supp } P$. ■

The following lemma is needed for the proof of [Proposition D.1.2](#).

Lemma D.1.4. *There is a positive constant $\sigma_1 > 0$ which depends on R and d such that, for $\sigma \in (0, \sigma_1]$ and $\xi \in \text{supp } P$,*

$$\left| \log \frac{Z(\xi, \sigma)}{(2\pi)^{d/2}} \right| \leq d|\log \sigma| + \log 2. \quad (\text{D.1.23})$$

Proof. It suffices to show that

$$\frac{(2\pi\sigma^2)^{d/2}}{2} \leq Z(\xi, \sigma) \leq (2\pi\sigma^2)^{d/2}, \quad (\text{D.1.24})$$

for any $\sigma \in (0, \sigma_1]$ with some $\sigma_1 > 0$ suitably defined. We prove the RHS by removing the constraint Ξ in the integral defining $Z(\xi, \sigma)$:

$$Z(\xi, \sigma) = \int_{\Xi} e^{-\frac{\|\xi - \zeta\|^2}{2\sigma^2}} d\zeta \leq \int_{\mathbb{R}^d} e^{-\frac{\|\xi - \zeta\|^2}{2\sigma^2}} d\zeta = (2\pi\sigma^2)^{d/2}. \quad (\text{D.1.25})$$

For the LHS, we invoke [Lemma D.7.1](#) using [Assumption 6.3](#) to get that $Z(\xi, \sigma) \geq \sigma^d$ when $\sigma \leq \sigma_1$ with $\sigma_1 > 0$ satisfying

$$1 - 6^{d/2} e^{-\frac{R^2}{12\sigma_1^2}} \geq \frac{1}{2}. \quad (\text{D.1.26})$$

■

We are now in a position to prove the main result of the section.

Proof of Proposition D.1.2. Applying [Lemmas D.1.3](#) and [D.1.4](#) and using the definition of G readily gives us that

$$|\phi(f, \xi, \lambda, \varepsilon, \sigma) - f(\xi)| \leq \frac{1}{2\lambda} G^2 + \frac{\varepsilon d}{2} \left| \log \left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2} \right) \right| + \varepsilon(d|\log \sigma| + \log 2) + \varepsilon \bar{Z}(\sigma) + \varepsilon c_1 e^{-c_2 \left(\frac{\lambda - L_2}{\varepsilon} \right)^{\frac{1}{3}}}. \quad (\text{D.1.27})$$

Since λ is always greater or equal than ε , $\log \left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2} \right)$ is always non-negative and, with λ belonging to $[\underline{\lambda}, \bar{\lambda}]$, we get that

$$|\phi(f, \xi, \lambda, \varepsilon, \sigma) - f(\xi)| \leq \frac{1}{2\underline{\lambda}} G^2 + \frac{\varepsilon d}{2} \log \left(\frac{\bar{\lambda}}{\varepsilon} + \frac{1}{\sigma^2} \right) + \varepsilon(d|\log \sigma| + \log 2) + \varepsilon c_1 e^{-c_2 \left(\frac{\bar{\lambda} - L_2}{\varepsilon} \right)^{\frac{1}{3}}}, \quad (\text{D.1.28})$$

which is the desired result. ■

As a consequence of this result, we have the following bound on the dual generator.

Corollary D.1.5. *For any $f \in \mathcal{F}$, $\xi \in \text{supp } P$, $\lambda \in [\underline{\lambda}, \bar{\lambda}]$, $\varepsilon \geq 0$ and $\sigma > 0$, the bound*

$$-a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \leq \phi(f, \xi, \lambda, \varepsilon, \sigma) \leq \tilde{F}(\underline{\lambda}), \quad (\text{D.1.29})$$

holds where

$$a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) := \begin{cases} 0 & \text{when } \varepsilon = 0 \\ M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) & \text{when } 0 < \varepsilon \leq \varepsilon_1, 0 < \sigma \leq \sigma_1, \underline{\lambda} \geq \lambda_1 + L_2 \\ G\sqrt{2C(\sigma)} + (L_2 + \bar{\lambda})C(\sigma) & \text{otherwise.} \end{cases} \quad (\text{D.1.30})$$

with $M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)$ the bounding term appearing in [Proposition D.1.2](#), as well as ε_1 , σ_1 , λ_1 .

Proof. For the upper-bound, it suffices to note that

$$\phi(f, \xi, \lambda, \varepsilon, \sigma) \leq \phi(f, \xi, \lambda, 0) \leq \phi(f, \xi, \underline{\lambda}, 0) \leq \tilde{F}(\underline{\lambda}) \quad (\text{D.1.31})$$

by definition of $\tilde{F}(\underline{\lambda})$. Let us now turn to the lower bound.

When $\varepsilon = 0$, we have that $\phi(f, \xi, \lambda, \varepsilon, \sigma) \geq f(\xi) \geq 0$.

When $0 < \varepsilon \leq \varepsilon_1$ and $\underline{\lambda} \geq \lambda_1 + L_2$, we have from [Proposition D.1.2](#)

$$\phi(f, \xi, \lambda, \varepsilon, \sigma) \geq f(\xi) - M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \geq -M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma). \quad (\text{D.1.32})$$

Otherwise, the bound comes from the smoothness of f and Jensen's inequality as

$$\phi(f, \xi, \lambda, \varepsilon, \sigma) \geq \varepsilon \log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \exp \left(\frac{f(\xi) + \langle \nabla f(\xi), \zeta - \xi \rangle - (L_2 + \lambda) \|\xi - \zeta\|^2/2}{\varepsilon} \right) \right) \quad (\text{D.1.33})$$

$$\geq \mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[f(\xi) + \langle \nabla f(\xi), \zeta - \xi \rangle - (L_2 + \lambda) \frac{1}{2} \|\xi - \zeta\|^2 \right] \quad (\text{D.1.34})$$

$$\geq - \left(G\sqrt{2C(\sigma)} + (L_2 + \lambda) C(\sigma) \right) \geq - \left(G\sqrt{2C(\sigma)} + (L_2 + \bar{\lambda}) C(\sigma) \right). \quad (\text{D.1.35})$$

■

D.1.4 Relative compactness of the class $\mathcal{F}$ of loss functions

In this section we prove the relative compactness of the class $\mathcal{F}$ w.r.t. several metrics. First, we show in [Lemma D.1.6](#) that, under our blanket assumptions, $\mathcal{F}$ is relatively compact for the infinity norm over Ξ , defined by i.e., $\|f\|_\infty := \sup_{\xi \in \Xi} |f(\xi)|$. Then, in [Lemma D.1.7](#), we establish the equivalence between the first item of [Assumption 6.5](#) and the relative compactness of $\mathcal{F}$ w.r.t. another distance that we introduce, as mentioned below [Assumption 6.5](#) in [§ 6.3.1](#). Finally, we leverage these compactness properties to ensure that the

Dudley integral of $\mathcal{F}$ w.r.t. those metrics, a standard complexity measure in concentration theory, is finite in [Lemma D.1.9](#).

Lemma D.1.6. $\mathcal{F}$ and $\text{conv}(\mathcal{F})$ are relatively compact for the topology of the uniform convergence.

Proof. First, the functions of $\mathcal{F}$ are uniformly Lipschitz-continuous: fix $\xi \in \text{supp } P$, then, for any $\zeta \in \Xi$ $\|\nabla f(\zeta)\| \leq L_2 \|\xi - \zeta\| + G \leq L_2 \sup_{(\xi, \zeta) \in \Xi} \|\xi - \zeta\| + G$ which is finite by compactness. Using the compactness of Ξ again, the functions in $\mathcal{F}$ are also uniformly bounded. As a consequence, the functions in $\text{conv}(\mathcal{F})$ are also uniformly Lipschitz-continuous and uniformly bounded. By the Arzelà-Ascoli theorem, see e.g., ([Rudin, 1987](#), Thm. 11.28), $\mathcal{F}$ and $\text{conv}(\mathcal{F})$ are then relatively compact for the topology of uniform convergence. ■

Recall that, for a set $A \subset \Xi$ and a point $\xi \in \Xi$, we denote by $d(\xi, A)$ the distance between ξ and A , i.e., $d(\xi, A) = \inf_{\zeta \in A} \|\xi - \zeta\|$.

Lemma D.1.7. Consider the distance, defined on continuous functions on Ξ by

$$D(f, g) := \|f - g\|_\infty + D_H(\arg \max f, \arg \max g) \quad (\text{D.1.36})$$

where D_H denotes the Hausdorff distance between sets associated to d , i.e., for $A, B \subset \Xi$,

$$D_H(A, B) := \max \left(\sup_{\xi \in A} d(\xi, B), \sup_{\xi \in B} d(\xi, A) \right). \quad (\text{D.1.37})$$

Under the blanket assumptions, we have that [Item 1](#) of [Assumption 6.5](#), i.e., that for any $R > 0$, there exists $\Delta > 0$ such that,

$$\forall f \in \mathcal{F}, \forall \zeta \in \Xi, d(\zeta, \arg \max f) \geq R \implies f(\zeta) - \max f \leq -\Delta, \quad (\text{D.1.38})$$

is equivalent to $\mathcal{F}$ being relatively compact for D .

Proof. $\implies$) Let us begin by showing that [\(D.1.38\)](#) implies the relative compactness of $\mathcal{F}$ for D , i.e., that the adherence of $\mathcal{F}$ is compact for D .

Take $(f_t)_{t=1,2,\dots}$ a sequence of functions from $\mathcal{F}$, and we will show that there is a subsequence which converges to some function in $\mathcal{F}$ for D . By compactness of $\mathcal{F}$ for the infinity norm, [Lemma D.1.6](#), there readily is a subsequence of $(f_t)_{t=1,2,\dots}$ that converges uniformly to some continuous function $f : \Xi \rightarrow \mathbb{R}$. Without loss of generality, let us assume that the whole sequence $(f_t)_{t=1,2,\dots}$ converges uniformly to f , i.e., that $\|f - f_t\|_\infty \rightarrow 0$ as $t \rightarrow +\infty$. As a consequence, it holds also holds that $\max_\Xi f_t$ converges to $\max_\Xi f$.

We now show that $D_H(\arg \max f_t, \arg \max f)$ converges to 0. $\mathcal{F}$ satisfy [\(D.1.38\)](#) by assumption. Hence, for any fixed $\eta > 0$, we can invoke [\(D.1.38\)](#) with $R \leftarrow \eta$ and it gives us some $\Delta > 0$. Now, since f is continuous, $\{\zeta \in \Xi : d(\zeta, \arg \max f) \geq \eta\}$ is a closed set inside a compact and therefore is compact as well. Hence, f reaches its maximum over this set and it is strictly less than $\max_\Xi f$ by construction. Substituting Δ with $\min(\Delta, \max_\Xi f - \max\{f(\zeta) : \zeta \in \Xi, d(\zeta, \arg \max f) \geq \eta\})$ which is still positive, we get that, for any $\zeta \in \Xi$, both,

$$d(\zeta, \arg \max f) \geq \eta \implies f(\zeta) - \max f \leq -\Delta, \quad (\text{D.1.39})$$

and, for any $t = 1, 2, \dots$,

$$d(\zeta, \arg \max f_t) \geq \eta \implies f_t(\zeta) - \max f_t \leq -\Delta. \quad (\text{D.1.40})$$

By convergence of the sequence, as mentioned above, there is some $T \geq 1$ such that, for any $t \geq T$, $\|f - f_t\|_\infty \leq \Delta/3$ and $|\max_{\Xi} f_t - \max_{\Xi} f| \leq \Delta/3$. These two inequalities imply that, for any $\xi \in \arg \max f$,

$$\max_{\Xi} f_t - f_t(\xi) \leq \max_{\Xi} f + \frac{\Delta}{3} - f(\xi) + \frac{\Delta}{3} = \frac{2\Delta}{3}. \quad (\text{D.1.41})$$

Therefore, by definition of Δ , it holds that $d(\xi, \arg \max f_t) < \eta$. Similarly, when $\xi \in \arg \max f_t$, one shows that $\max_{\Xi} f - f(\xi) \leq 2\Delta/3$ so that we have $d(\xi, \arg \max f) < \eta$ as well. Hence, for any $t \geq T$, $D_H(\arg \max f_t, \arg \max f)$ is at most η .

Therefore, we have shown that $D_H(\arg \max f_t, \arg \max f)$ goes to zero. Since $\|f - f_t\|_\infty$ converges to zero as well by construction, this means that $D(f_t, f)$ converges to zero, which concludes the proof.

($\Leftarrow$) Let us proceed by contradiction, i.e., assume that there is some $R > 0$, some sequence $(f_t)_{t=1,2,\dots}$ of functions from $\mathcal{F}$ and some sequence $(\xi_t)_{t=1,2,\dots}$ of points from Ξ such that,

$$\forall t = 1, 2, \dots, d(\arg \max f_t, \xi_t) \geq R \quad \text{yet} \quad f_t(\xi_t) - \max_{\Xi} f_t \rightarrow 0 \text{ as } t \rightarrow +\infty. \quad (\text{D.1.42})$$

Since Ξ is compact and since we assume $\mathcal{F}$ to be relatively compact for D , without loss of generality, we can assume that $(\xi_t)_{t=1,2,\dots}$ converges to some $\xi \in \Xi$ while $(f_t)_{t=1,2,\dots}$ converges to some continuous function f for D . On the one hand, by definition of the Hausdorff distance, we have that, for any $t = 1, 2, \dots$,

$$d(\arg \max f, \xi) \geq d(\arg \max f_t, \xi) - D_H(\arg \max f_t, \arg \max f) \quad (\text{D.1.43})$$

$$\geq d(\arg \max f_t, \xi_t) - (d(\xi, \xi_t) + D_H(\arg \max f_t, \arg \max f)), \quad (\text{D.1.44})$$

so that, by taking $t \rightarrow +\infty$, we get that $d(\arg \max f, \xi) \geq R$. On the other hand, by uniform convergence, one has that

$$f(\xi) - \max f = \lim_{t \rightarrow +\infty} f_t(\xi_t) - \max f_t = 0, \quad (\text{D.1.45})$$

which yields the contradiction since ξ cannot belong to $\arg \max f$. ■

Note that, for parametric models (§ 6.4.5), this lemma gives a computation-free approach to verifying the second item of [Assumption 6.5](#).

Corollary D.1.8. *Consider Θ a compact subset of $\mathbb{R}^p$ and $f : \Theta \times \Xi \rightarrow \mathbb{R}$ a continuous function. If the map $\theta \in \Theta \mapsto f(\theta, \cdot)$ is continuous from Θ to the space of continuous functions on Ξ equipped with the distance D defined in [Lemma D.1.7](#), then $\mathcal{F} := \{f(\theta, \cdot) : \theta \in \Theta\}$ is compact for D .*

In particular, this corollary allows one to easily check that [Examples 6.1](#) and [6.2](#) satisfy the second item of [Assumption 6.5](#).

We finally introduce Dudley's integral, which is a standard complexity measure in concentration theory.

Definition D.2. Dudley's entropy integral $\mathcal{I}(\mathcal{X}, \text{dist})$ is defined for a metric space $(\mathcal{X}, \text{dist})$ as

$$\mathcal{I}(\mathcal{X}, \text{dist}) := \int_0^{+\infty} \sqrt{\log N(t, \mathcal{X}, \text{dist})} dt \quad (\text{D.1.46})$$

where $N(t, \mathcal{X}, \text{dist})$ denotes the t -packing number of $\mathcal{X}$, which is the maximal number of points in $\mathcal{X}$ which are at least at a distance t from each other.

Lemma D.1.9. *The Dudley integral of $\mathcal{F}$ w.r.t. $\|\cdot\|_\infty$, that we denote by $\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)$, is finite. Under [Assumption 6.5](#), the Dudley integral of $\mathcal{F}$ w.r.t. D , denoted by $\mathcal{I}(\mathcal{F}, D)$ is finite as well.*

Proof. [Lemma D.1.6](#) shows that $\mathcal{F}$ is relatively compact for the norm $\|\cdot\|_\infty$ and in particular bounded. Since Dudley's entropy integral is finite for balls ([Wainwright, 2019](#), Ex. 5.18) and $\mathcal{F}$ is now included in some ball for $\|\cdot\|_\infty$, the integral $\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)$ is indeed finite. The second assertion is proven using the same reasoning and [Lemma D.1.7](#). ■

D.1.5 Parametric Morse-Bott objectives

In this section, we discuss the quadratic growth condition of the second item of [Assumption 6.5](#) and its relation to the parametric Morse-Bott assumption of [Arbel and Mairal \(2022\)](#). Indeed, in the context of smooth manifolds and parametric models, we prove that the parametric Morse-Bott assumption implies the quadratic growth condition of [Assumption 6.5](#). In [Assumption D.3](#), we introduce the Riemannian and parametric settings that are necessary to formulate the parametric Morse-Bott condition and we then present a version of this condition adapted to our context. We refer to [Lee \(2018\)](#) for definitions relevant to Riemannian geometry. The main result of this section is then [Proposition D.1.10](#), which relies on [Lemma D.1.11](#) for its proof.

Assumption D.3 (Parametric Morse-Bott). Let $\mathcal{F} = \{\xi \mapsto f(\theta, \xi) : \theta \in \Theta\}$ where :

- Ξ, Θ are smooth compact (connected embedded) submanifolds of $\mathbb{R}^d$ and $\mathbb{R}^p$ respectively, endowed with the induced Euclidean metric.
- $f : \Theta \times \Xi \rightarrow \mathbb{R}$ is thrice continuously differentiable on the product manifold.
- f is a parametric Morse-Bott function ([Arbel and Mairal, 2022](#), Def. 2): the set of augmented critical points of $\mathcal{F}$, defined as

$$M := \{(\theta, \xi) \in \Theta \times \Xi : \text{grad}_\xi f(\theta, \xi) = 0\}. \quad (\text{D.1.47})$$

is a smooth (embedded) submanifold of $\Theta \times \Xi \setminus \text{bd } \Xi$ whose dimension at $(\theta, \xi) \in M$ is $\dim_\theta(\Theta) + \dim(\ker \text{Hess}_\xi f(\theta, \xi))$.

Under this assumption [Assumption D.3](#), the following result thus guarantees that the quadratic growth condition of [Assumption 6.5](#) holds.

Proposition D.1.10. *Under [Assumption D.3](#) and the first item of [Assumption 6.5](#), the second item of [Assumption 6.5](#) holds, i.e., there exists $\mu, L_3 > 0$ such that, for all $\theta \in \Theta$, $\xi \in \Xi$ and $\xi^* \in \arg \max f$ a projection of ξ on $\arg \max f$, i.e., $\xi^* \in \arg \min_{\arg \max f} \|\xi - \cdot\|$, it holds that*

$$f(\theta, \xi^*) \geq f(\theta, \xi) + \frac{\mu}{2} \|\xi - \xi^*\|^2 - \frac{L_3}{6} \|\xi - \xi^*\|^3. \quad (\text{D.1.48})$$

To show this result, we rely on the following lemma that relates [Assumption D.3](#) to a local quadratic growth condition.

Lemma D.1.11. *Under [Assumption D.3](#), for any $(\theta_0, \xi_0) \in M$ such that ξ_0 is a local maximum of $f(\theta_0, \cdot)$ and any neighborhood $\mathcal{W}$ of (θ_0, ξ_0) in M , there exists a neighborhood $\mathcal{U}$ of (θ_0, ξ_0) in $\Theta \times \Xi$ and $\mu > 0$ such that, for any $(\theta, \xi) \in \mathcal{U}$, there exists $\xi^* \in \Xi$ such that $(\theta, \xi^*) \in \mathcal{W}$ and*

$$f(\theta, \xi^*) \geq f(\theta, \xi) + \frac{\mu}{2} \|\xi - \xi^*\|^2. \quad (\text{D.1.49})$$

Proof. By assumption, the tangent space of M at (θ, ξ) is given by

$$T_{(\theta, \xi)} M = T_\theta \Theta \times \ker \text{Hess}_\xi f(\theta, \xi), \quad (\text{D.1.50})$$

and so its normal space (in $\Theta \times \Xi$) is equal to

$$N_{(\theta, \xi)} M = \{0\} \times (\ker \text{Hess}_\xi f(\theta, \xi))^\perp \subset T_\theta \Theta \times T_\xi \Xi. \quad (\text{D.1.51})$$

Applying the inverse function theorem to the normal exponential map $E : (\theta, \xi, (0, z)) \in NM \mapsto (\theta, \exp_\xi(z))$ following the proof of [Lee \(2018, Thm. 5.25\)](#), there exists $\eta > 0$ and a neighborhood $\mathcal{U}_\eta$ of M in $\Theta \times \Xi$ such that, with

$$\mathcal{V}_\eta := \{(\theta, \xi, z) : (\theta, \xi) \in M, z \in (\ker \text{Hess}_\xi f(\theta, \xi))^\perp, \|z\| < \eta\}, \quad (\text{D.1.52})$$

the normal exponential map E is a diffeomorphism from $\mathcal{V}_\eta$ to $\mathcal{U}_\eta$. Note that $\mathcal{V}_\eta$ is relatively compact and, as a consequence, the third derivative of $t \in [0, 1] \mapsto f(\theta, \exp_\xi(tz))$ is a continuous function of $t \in [0, 1]$ and $(\theta, \xi, z) \in \mathcal{V}_\eta$ and as a consequence is bounded uniformly by some constant $L_3 > 0$. Fix $(\theta_0, \xi_0) \in M$ such that ξ_0 is a local maximum of $f(\theta, \cdot)$. Consider the map

$$v : (\theta, \xi) \mapsto \inf \{ \langle z, -\text{Hess}_\xi f(\theta, \xi)z \rangle : z \in (\ker \text{Hess}_\xi f(\theta, \xi))^\perp, \|z\| = 1 \}. \quad (\text{D.1.53})$$

If $v(\theta_0, \xi_0)$ is $+\infty$, i.e., if $\ker \text{Hess}_\xi f(\theta_0, \xi_0)$ is equal to the whole $T_{\xi_0} \Xi$, then, since the dimension of a manifold is locally constant, there is a neighborhood of (θ_0, ξ_0) in $\Theta \times \Xi$ on which v is identically equal to $+\infty$. Otherwise, if $v(\theta_0, \xi_0)$ is finite, then it is positive by construction. Hence, the continuity of v implies there is a positive constant $\mu > 0$ and a neighborhood of (θ_0, ξ_0) in $\Theta \times \Xi$ on which v is lower-bounded by μ .

Hence, in both cases, there is $\mu > 0$ and $\mathcal{V}'$ a neighborhood of (θ_0, ξ_0) in $\Theta \times \Xi$ such that v is at least greater or equal to μ on $\mathcal{V}'$. Finally, take

$$\mathcal{U} := \mathcal{U}_\eta \cap E \left(\left\{ (\theta, \xi, z) \in \mathcal{V}_\eta : (\theta, \xi) \in \mathcal{V}', \|z\| < \frac{3\mu}{4L_3} \right\} \right). \quad (\text{D.1.54})$$

We are now in a position to prove the result. Take $(\theta, \xi) \in \mathcal{U}$. Since $\mathcal{U}$ is included in $\mathcal{U}_\eta$, there is some $\xi^* \in \Xi$ and $z \in (\ker \text{Hess}_\xi f(\theta, \xi))^\perp$ such that $(\theta, \xi^*) \in M \cap \mathcal{V}' \cap \mathcal{W}$, $\|z\| < \frac{3\mu}{4L_3}$ and $\exp_\xi^*(z) = \xi^*$. Let $\gamma(t) := \exp_\xi^*(tz)$ for $t \in [0, 1]$ denotes the geodesic curve going from ξ^* to ξ . Then, by the Taylor inequality applied to $t \mapsto f(\xi, \gamma(t))$ (see [Boumal \(2023, § 5.9\)](#)) and by definition of L_3 ,

$$f(\theta, \xi) \leq f(\theta, \xi^*) + \langle \text{grad}_\xi f(\theta, \xi^*), z \rangle + \frac{1}{2} \langle \text{Hess}_\xi f(\theta, \xi^*)z, z \rangle \quad (\text{D.1.55})$$

$$+ \frac{1}{2} \langle \text{grad}_\xi f(\theta, \xi^*), \gamma''(0) \rangle + \frac{L_3}{6} \|z\|^3. \quad (\text{D.1.56})$$

But $\gamma''(t)$ is null since γ is a geodesic and $\text{grad}_\xi f(\theta, \xi^*)$ too by definition. Moreover, since $(\theta, \xi^*) \in \mathcal{V}$ and $z \in (\ker \text{Hess}_\xi f(\theta, \xi^*))^\perp$, the term $\langle \text{Hess}_\xi f(\theta, \xi^*)z, z \rangle$ is bounded by $-\mu\|z\|^2$. But $\|z\|$ is also equal to $\|\xi - \xi^*\|$ by definition of z so we get,

$$f(\theta, \xi) \leq f(\theta, \xi^*) - \frac{\mu}{2}\|\xi - \xi^*\|^2 + \frac{L_3}{6}\|\xi - \xi^*\|^3 \quad (\text{D.1.57})$$

$$\leq f(\theta, \xi^*) - \frac{\mu}{4}\|\xi - \xi^*\|^2, \quad (\text{D.1.58})$$

since $\|z\| = \|\xi - \xi^*\| \leq \frac{3\mu}{4L_3}$, which gives the result. $\blacksquare$

We are now ready to prove [Proposition D.1.10](#).

Proof of Proposition D.1.10. We build upon the result of [Lemma D.1.11](#). Fix $(\theta_0, \xi_0) \in M$ such that ξ_0 is a maximum of $f(\theta_0, \cdot)$ and let $r > 0$ such that $\overline{\mathbb{B}}((\theta_0, \xi_0), r) \cap M$ is diffeomorphic to an Euclidean ball. Invoke the first item of [Assumption 6.5](#), with $R \leftarrow r/2$ and let $\Delta > 0$ be the given positive quantity. Let $\mathcal{U}, \mu$ be given by [Lemma D.1.11](#) invoked with $\mathcal{W} := \overline{\mathbb{B}}((\theta_0, \xi_0), \frac{r}{2}) \cap \{(\theta, \xi) \in M : f(\theta, \xi) > \max_{\Xi} f(\theta, \cdot) - \Delta\}$.

Hence, for any $(\theta, \xi) \in \mathcal{U}$, there is $\xi^* \in \Xi$ such that $(\theta, \xi^*) \in M \cap \mathcal{W}$ and

$$f(\theta, \xi^*) \geq f(\theta, \xi) + \frac{\mu}{2}\|\xi - \xi^*\|^2. \quad (\text{D.1.59})$$

But (θ, ξ^*) also satisfies $f(\theta, \xi^*) > \max_{\Xi} f(\theta, \cdot) - \Delta$ so that $d(\xi^*, \arg \max_{\Xi} f(\theta, \cdot)) < \frac{r}{2}$ by definition of Δ , i.e., there exists ξ^{**} that is a maximizer of $f(\theta, \cdot)$ and that is at distance at most $< \frac{r}{2}$ from ξ^* . But then both ξ^* and ξ^{**} belong to $\overline{\mathbb{B}}((\theta_0, \xi_0), r)$ that is diffeomorphic to an Euclidean ball. Hence, since the derivative of $f(\theta, \cdot)$ is null on M , $f(\theta, \xi^*) = f(\theta, \xi^{**}) = \max_{\Xi} f(\theta, \cdot)$ so that ξ^* is a maximizer of $f(\theta, \cdot)$ too. Therefore, (D.1.59) becomes

$$\max_{\Xi} f(\theta, \cdot) = f(\theta, \xi^*) \geq f(\theta, \xi) + \frac{\mu}{2}\|\xi - \xi^*\|^2 \geq f(\theta, \xi) + \frac{\mu}{2}d^2\left(\xi, \arg \max_{\Xi} f(\theta, \cdot)\right). \quad (\text{D.1.60})$$

The final statement of the proposition follows by compactness and uniform Lipschitz-continuity of $\mathcal{F}$ (see the proof of [Lemma D.1.6](#)). $\blacksquare$

D.2 FROM EMPIRICAL TO TRUE RISK VIA DUALITY

In this part of the proof, our objective is to show that: if the dual variable λ in (6.19) can be bounded uniformly in $[\underline{\lambda}, \bar{\lambda}]$ with probability $1 - \delta$, then we can concentrate the empirical expectation in (6.19) towards the one in (6.20). The concentration error induces a loss in the radius, fortunately, captured by the variable $\bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)$ that we take as

$$\bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) := \frac{117}{\sqrt{n\underline{\lambda}}} \left(\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \max\left(\tilde{F}(\underline{\lambda}), a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)\right) \left(1 + \sqrt{\log \frac{1}{\delta}}\right) \right), \quad (\text{D.2.1})$$

where $\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)$ is the Dudley integral of $\mathcal{F}$ w.r.t. the infinity norm (Definition D.2), $\tilde{F}$ is defined in Assumption D.1, and $a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)$ is the bounding term appearing in Corollary D.1.5.

The main result of this part is Proposition D.2.1, stated below, and the remainder of the section will consist in proving it.

Proposition D.2.1. *for $\rho > 0$, $\varepsilon \geq 0$, $\sigma > 0$ and $\delta \in (0, 1)$, assume that there is some $0 < \underline{\lambda} \leq \bar{\lambda} < +\infty$ such that, with probability at least $1 - \frac{\delta}{2}$,*

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] . \quad (\text{D.2.2})$$

then, when $\rho^2 \geq \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)$, with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \mathcal{R}_{\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)}^\varepsilon(f) . \quad (\text{D.2.3})$$

The proof of this result mainly consists in verifying that under our standing assumptions, we can apply the concentration result presented in Lemma D.7.2 in order to concentrate $\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f)$ towards $\mathcal{R}_{\rho^2}^\varepsilon(f)$ through their dual formulations.

We begin by showing that the dual generator *divided by λ* is Lipschitz continuous in f and in λ^{-1} (for convenience, we use the notation $\mu = \lambda^{-1}$).

Lemma D.2.2. *Fix some $\bar{\lambda} \geq \underline{\lambda} > 0$. For any $\xi \in \Xi$, $\varepsilon \geq 0$ and $\sigma > 0$ we have that*

- (a) *for any $\lambda \in [\underline{\lambda}, \bar{\lambda}]$, $f \mapsto \lambda^{-1} \phi(f, \xi, \lambda, \varepsilon, \sigma)$ is $\underline{\lambda}^{-1}$ -Lipschitz continuous w.r.t. the norm $\|\cdot\|_\infty$;*
- (b) *for any $f \in \mathcal{F}$, $\mu \mapsto \mu \phi(f, \xi, \mu^{-1}, \varepsilon, \sigma)$ is $\max(\tilde{F}(\underline{\lambda}), a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma))$ -Lipschitz continuous on $[\bar{\lambda}^{-1}, \underline{\lambda}^{-1}]$.*

Proof. Item (a). When $\varepsilon = 0$, $f \mapsto \phi(f, \xi, \lambda, \varepsilon, \sigma)$ is a supremum of 1-Lipschitz functions and is thus 1-Lipschitz. For $\varepsilon > 0$, take $f, g \in \mathcal{F}$ and, for $t \in [0, 1]$, define $f_t = f + t(g - f)$. Differentiating $t \mapsto \phi(f_t, \xi, \lambda, \varepsilon, \sigma)$ yields

$$\left| \frac{d}{dt} \phi(f_t, \xi, \lambda, \varepsilon, \sigma) \right| = \left| \frac{\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[(g(\zeta) - f(\zeta)) e^{\frac{f_t(\zeta) - \lambda \|\xi - \zeta\|^2/2}{\varepsilon}} \right]}{\mathbb{E}_{\zeta' \sim \pi_\sigma(\cdot|\xi)} \left[e^{\frac{f_t(\zeta') - \lambda \|\xi - \zeta'\|^2/2}{\varepsilon}} \right]} \right| \leq \|g - f\|_\infty , \quad (\text{D.2.4})$$

which gives that $f \mapsto \phi(f, \xi, \lambda, \varepsilon, \sigma)$ is 1-Lipschitz continuous w.r.t. the norm $\|\cdot\|_\infty$.

Since this bound is uniform in λ , we immediately get that $f \mapsto \lambda^{-1} \phi(f, \xi, \lambda, \varepsilon, \sigma)$ is $\underline{\lambda}^{-1}$ -Lipschitz continuous for all $\lambda \geq \underline{\lambda}$.

Item (b). Fix $f \in \mathcal{F}$, $\xi \in \Xi$, $\varepsilon \geq 0$, $\sigma > 0$ and define $g(\lambda) := \lambda^{-1} \phi(f, \xi, \lambda, \varepsilon, \sigma)$.

Let us first begin with the case $\varepsilon = 0$. Take $\lambda, \lambda' \in [\underline{\lambda}, \bar{\lambda}]$. Without loss of generality, we can suppose that $g(\lambda) \geq g(\lambda')$. Since f is continuous and Ξ is a compact set, choose $\zeta \in \arg \max_{\zeta \in \Xi} \{f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2\}$. Then, the claim comes from the fact that

$$0 \leq g(\lambda) - g(\lambda') \leq \lambda^{-1} f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2 - \left(\lambda'^{-1} f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2 \right) \leq |\lambda^{-1} - \lambda'^{-1}| \tilde{F}(\underline{\lambda}) , \quad (\text{D.2.5})$$

where we use that since f is non-negative by assumption, $|\widetilde{F}(\lambda)| = \widetilde{F}(\lambda) \leq \widetilde{F}(\underline{\lambda})$.

Let us now turn to the case where $\varepsilon > 0$, for which g is differentiable on $[\underline{\lambda}, \bar{\lambda}]$ with derivative

$$g'(\lambda) = -\frac{1}{\lambda^2} \phi(f, \xi, \lambda, \varepsilon, \sigma) - \frac{1}{\lambda} \frac{\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 e^{\frac{f(\zeta) - \lambda \|\xi - \zeta\|^2/2}{\varepsilon}} \right]}{\mathbb{E}_{\zeta' \sim \pi_\sigma(\cdot|\xi)} \left[e^{\frac{f(\zeta') - \lambda \|\xi - \zeta'\|^2/2}{\varepsilon}} \right]}. \quad (\text{D.2.6})$$

Since the claimed result is the Lipschitz continuity of $h : \mu \mapsto g(\mu^{-1})$, it suffices to bound its derivative, i.e., to bound $-\lambda^{-2} g'(\lambda)$ for all $\bar{\lambda} \geq \lambda \geq \underline{\lambda}$. On the one hand, thanks to [Lemma D.7.7](#), it is bounded above as

$$-\frac{1}{\lambda^2} g'(\lambda) \leq \frac{\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[(f(\zeta) - \frac{\lambda}{2} \|\xi - \zeta\|^2) e^{\frac{f(\zeta) - \lambda \|\xi - \zeta\|^2/2}{\varepsilon}} \right]}{\mathbb{E}_{\zeta' \sim \pi_\sigma(\cdot|\xi)} \left[e^{\frac{f(\zeta') - \lambda \|\xi - \zeta'\|^2/2}{\varepsilon}} \right]} + \lambda \frac{\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 e^{\frac{f(\zeta) - \lambda \|\xi - \zeta\|^2/2}{\varepsilon}} \right]}{\mathbb{E}_{\zeta' \sim \pi_\sigma(\cdot|\xi)} \left[e^{\frac{f(\zeta') - \lambda \|\xi - \zeta'\|^2/2}{\varepsilon}} \right]} \quad (\text{D.2.7})$$

$$\leq \phi(f, \xi, \lambda, 0) \leq \widetilde{F}(\lambda) \leq \widetilde{F}(\underline{\lambda}). \quad (\text{D.2.8})$$

On the other hand, invoking [Corollary D.1.5](#) also yields that

$$-\frac{1}{\lambda^2} g'(\lambda) \geq \varepsilon \log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} e^{\frac{f(\zeta) - \lambda \|\xi - \zeta\|^2/2}{\varepsilon}} \right) \geq -a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma), \quad (\text{D.2.9})$$

which concludes the proof. $\blacksquare$

We can now apply standard concentration for bounded Lipschitz quantities to bound the difference between the expectation of the dual generator over the empirical distribution $\widehat{\mathbb{P}}_n$ and true one $\mathbb{P}$.

Lemma D.2.3. *For $\rho > 0$, $\varepsilon \geq 0$, $\sigma > 0$, $\delta \in (0, 1)$ and some $0 < \underline{\lambda} \leq \bar{\lambda} < +\infty$, we have with probability at least $1 - \frac{\delta}{2}$ that*

$$\sup_{(f, \lambda) \in \mathcal{F} \times [\underline{\lambda}, \bar{\lambda}]} \left\{ \frac{\mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]}{\lambda} \right\} \leq \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma). \quad (\text{D.2.10})$$

Proof. Our objective is to bound the quantity

$$\sup_{(f, \lambda) \in \mathcal{F} \times [\underline{\lambda}, \bar{\lambda}]} \left\{ \frac{\mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \widehat{\mathbb{P}}_n, \lambda, \varepsilon, \sigma)]}{\lambda} \right\} \quad (\text{D.2.11})$$

$$= \sup_{(f, \mu) \in \mathcal{F} \times [\bar{\lambda}^{-1}, \underline{\lambda}^{-1}]} \left\{ \mathbb{E}_{\xi \sim \mathbb{P}} [\mu \phi(f, \xi, \mu^{-1}, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\mu \phi(f, \xi, \mu^{-1}, \varepsilon, \sigma)] \right\} \quad (\text{D.2.12})$$

$$= \sup_{(f, \mu) \in \mathcal{X}} \left\{ \mathbb{E}_{\xi \sim \mathbb{P}} [X((f, \mu), \xi)] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [X((f, \mu), \xi)] \right\}, \quad (\text{D.2.13})$$

where we used again the notation $\mu = \lambda^{-1}$ and defined

$$\mathcal{X} := \mathcal{F} \times \left[\bar{\lambda}^{-1}, \underline{\lambda}^{-1} \right] \quad \text{and} \quad X((f, \mu), \xi) := \mu \phi(f, \xi, \mu^{-1}, \varepsilon, \sigma). \quad (\text{D.2.14})$$

Let us endow $\mathcal{X}$ with the distance,

$$\text{dist}((f, \mu), (f', \mu')) := \underline{\lambda}^{-1} \|f - f'\|_{\infty} + \max\left(\tilde{F}(\underline{\lambda}), a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)\right) |\mu - \mu'|. \quad (\text{D.2.15})$$

We now wish to apply [Lemma D.7.2](#) and check its three requirements:

1. For any $(f, \mu) \in \mathcal{F} \times \left[\bar{\lambda}^{-1}, \underline{\lambda}^{-1} \right]$, $X((f, \mu), \cdot)$ is measurable since the functions of $\mathcal{F}$ are continuous and thus *a fortiori* measurable;
2. By [Lemma D.2.2](#), for any $\varepsilon \geq 0$ and any $\xi \in \text{supp } P$, $X(\cdot, \xi)$ is 1-Lipschitz w.r.t. dist ;
3. Thanks to [Corollary D.1.5](#), for any $(f, \mu) \in \mathcal{X}$, $\xi \in \text{supp } P$, $\varepsilon \geq 0$ and $\sigma > 0$, we have

$$-\frac{a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)}{\underline{\lambda}} \leq X((f, \mu), \xi) \leq \frac{\tilde{F}(\underline{\lambda})}{\underline{\lambda}}. \quad (\text{D.2.16})$$

As a consequence, applying statement (b) of [Lemma D.7.2](#) yields that, with probability at least $1 - \frac{\delta}{2}$,

$$\sup_{(f, \lambda) \in \mathcal{F} \times [\underline{\lambda}, \bar{\lambda}]} \left\{ \frac{\mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim \hat{P}_n} [\phi(f, \hat{P}_n, \lambda, \varepsilon, \sigma)]}{\lambda} \right\} \quad (\text{D.2.17})$$

$$\leq \frac{48\mathcal{I}(\mathcal{X}, \text{dist})}{\sqrt{n}} + \frac{2}{\underline{\lambda}} \left(\tilde{F}(\underline{\lambda}) + a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \right) \sqrt{\frac{\log \frac{2}{\delta}}{2n}}. \quad (\text{D.2.18})$$

We now proceed to bound $\mathcal{I}(\mathcal{X}, \text{dist})$. Exploiting the product space structure of $\mathcal{X}$ and dist with [Lemma D.7.3](#), one has that,

$$\mathcal{I}(\mathcal{X}, \text{dist}) \leq \underline{\lambda}^{-1} \mathcal{I}(\mathcal{F}, \|\cdot\|_{\infty}) + \max\left(\tilde{F}(\underline{\lambda}), a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)\right) \mathcal{I}([0, \underline{\lambda}^{-1}], |\cdot|) \quad (\text{D.2.19})$$

$$\leq \underline{\lambda}^{-1} \left(\mathcal{I}(\mathcal{F}, \|\cdot\|_{\infty}) + \max\left(\tilde{F}(\underline{\lambda}), a(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)\right) \frac{1 + 2 \log 2}{2} \right), \quad (\text{D.2.20})$$

where we used [Lemma D.7.4](#). Hence, we have shown that with probability at least $1 - \frac{\delta}{2}$,

$$\sup_{(f, \lambda) \in \mathcal{F} \times [\underline{\lambda}, \bar{\lambda}]} \left\{ \frac{\mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim \hat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]}{\lambda} \right\} \leq \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \quad (\text{D.2.21})$$

where some numerical constants have been simplified. ■

Proof of Proposition D.2.1. Building on Lemma D.2.3, we can now conclude the main result of this section. Using our boundedness assumption on λ , we have that, with probability $1 - \delta$, the two following statements hold simultaneously

$$\bullet \quad \forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] ; \quad (\text{D.2.22})$$

$$\bullet \quad \sup_{(f, \lambda) \in \mathcal{F} \times [\underline{\lambda}, \bar{\lambda}]} \left\{ \frac{\mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]}{\lambda} \right\} \leq \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma). \quad (\text{D.2.23})$$

As a consequence, on this event, for any $f \in \mathcal{F}$,

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \right\} \quad (\text{D.2.24})$$

$$= \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \lambda \frac{\mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]}{\lambda} \right\} \quad (\text{D.2.25})$$

$$\geq \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \lambda \sup_{\underline{\lambda} \leq \lambda' \leq \bar{\lambda}} \frac{\mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda', \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda', \varepsilon, \sigma)]}{\lambda'} \right\} \quad (\text{D.2.26})$$

$$\geq \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \lambda \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \right\} \quad (\text{D.2.27})$$

$$\geq \mathcal{R}_{\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)}^\varepsilon(f) \quad (\text{D.2.28})$$

where $\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \geq 0$ by assumption. $\blacksquare$

Remark 12. Note that the proof of Proposition D.2.1 actually gives us the slightly stronger result at the penultimate equation: with probability at least $1 - \delta$, for any $f \in \mathcal{F}$,

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda(\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)) + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \right\}, \quad (\text{D.2.29})$$

that we will require later.

D.3 DUAL BOUND WHEN ρ IS SMALL

In this section, we show how the condition (D.2.2) of Proposition D.2.1 can be obtained when the robustness radius ρ is small enough. The results of this section cover both the standard WDRO setting of Theorems 6.1 and 6.2 and the regularized case of Theorem 6.6.

In the following Assumption D.4, we precise how small ρ has to be; we also take ε and σ proportional to ρ in order to get close to the true risk with ρ , ε and σ “small” at the same time. The main result of this section is Proposition D.3.1, whose proof relies on Lemma D.3.2.

Assumption D.4 (ρ is small). Take $\varepsilon = \varepsilon_0 \rho$, $\sigma = \sigma_0 \rho$ with $\varepsilon_0 \geq 0$, $\sigma_0 > 0$ and define

$$\lambda_0^* := \varepsilon_0 d + \sqrt{(\varepsilon_0 d)^2 + 8 \inf_{f \in \mathcal{F}} \mathbb{E}_P \|\nabla f\|_2^2} \quad \mu^* := \frac{8 \inf_{f \in \mathcal{F}} \mathbb{E}_P \|\nabla f\|_2^2}{(\lambda_0^*)^3} + \frac{2\varepsilon_0 d}{(\lambda_0^*)^2}. \quad (\text{D.3.1})$$

Moreover, assume that ε_0 and σ_0 satisfy

$$\frac{\varepsilon_0}{\sigma_0^2} \leq \frac{\lambda_0^*}{8}. \quad (\text{D.3.2})$$

Assume that $\rho > 0$ is small enough so that,

$$\rho \leq \min \left(\frac{\varepsilon_1}{\varepsilon_0}, \frac{\lambda_0^*}{32(\lambda_1 + L_2)}, \frac{\mu^*(\lambda_0^*)^2}{4096L_2}, \sqrt{\frac{c_2^3 \lambda_0^*}{8\varepsilon_0}} \left(\log \left(\frac{4096\varepsilon_0 c_1}{\mu^*(\lambda_0^*)^2} \right) \right)_+^{-\frac{3}{2}} \right) \quad (\text{D.3.3})$$

where $\lambda_1, \varepsilon_1, c_1, c_2$ are positive constant given by [Lemma D.1.3](#) and σ comes from [\(6.30\)](#).

Note that λ_0^* and μ^* are both always positive, be it thanks to [Assumption 6.5](#) or the regularization with $\varepsilon_0 > 0$. For such values of ρ , the main result of this section [Proposition D.3.1](#) shows that the dual variable of [\(6.19\)](#) can be bounded with high probability.

Proposition D.3.1. *Let [Assumption D.4](#) hold and fix a threshold $\delta \in (0, 1)$. Assume in addition that*

$$\rho \geq \frac{8192}{\sqrt{n} \mu^*(\lambda_0^*)^2} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + (\tilde{F}(0) + \overline{M}(\rho)) \sqrt{1 + \log \frac{1}{\delta}} \right) \quad (\text{D.3.4})$$

where $\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)$, $\tilde{F}$ are defined in [Appendix D.1.1](#) and $M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)$, the bounding term appearing in [Proposition D.1.2](#), is used to define

$$\overline{M}(\rho) := \sup_{\rho' \in (0, \rho]} M \left(\max \left(\frac{\lambda_0^*}{32\rho'}, \lambda_1 + L_2 \right), \frac{\lambda_0^*}{2\rho'}, \varepsilon_0 \rho', \sigma_0 \rho' \right). \quad (\text{D.3.5})$$

Then, with probability at least $1 - \delta$, we have

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\frac{\lambda_0^*}{32\rho} \leq \lambda \leq \frac{\lambda_0^*}{2\rho}} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]. \quad (\text{D.3.6})$$

To show [Proposition D.3.1](#), we need the following helper lemma.

Lemma D.3.2. *Let [Assumption D.4](#) hold. Then,*

$$\left(\frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0 \rho}{\sigma^2} + L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim P} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0 \rho}{\sigma^2} + L_2, \varepsilon \right) \right] + \frac{\rho \mu^*}{1024} (\lambda_0^*)^2 \quad (\text{D.3.7})$$

$$\leq \min \left(\left(\frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0 \rho}{2\sigma^2} - L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim P} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0 \rho}{2\sigma^2} - L_2, \varepsilon \right) \right], \quad (\text{D.3.8}) \right.$$

$$\left. \left(\frac{\lambda_0^*}{2\rho} - \frac{2\varepsilon_0 \rho}{\sigma^2} + L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim P} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{2\rho} - \frac{2\varepsilon_0 \rho}{\sigma^2} + L_2, \varepsilon \right) \right] \right) \quad (\text{D.3.9})$$

$$\text{and} \quad \max \left(\frac{\lambda_0^*}{32\rho}, \lambda_1 + L_2 \right) \leq \frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2 \leq \frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2 \leq \frac{\lambda_0^*}{2\rho} \quad (\text{D.3.10})$$

Proof. Fix $f \in \mathcal{F}$. Consider the function $\overline{\psi}_\rho : \lambda \mapsto \lambda\rho^2 + \mathbb{E}_{\xi \sim P} [\overline{\phi}(f, \xi, \lambda, \varepsilon, \sigma)]$ where $\overline{\phi}$ is defined in (D.1.12). By Lemma D.7.8 invoked with $a \leftarrow \rho^2$, $b \leftarrow \frac{1}{2}\mathbb{E}_P[\|\nabla f\|_2^2]$, $c \leftarrow \frac{\varepsilon d}{2}$ and $r \leftarrow \frac{\varepsilon}{\sigma^2}$, its unique minimizer is

$$\lambda^* := \left[\frac{\varepsilon_0 d + \sqrt{(\varepsilon_0 d)^2 + 8\mathbb{E}_P[\|\nabla f\|_2^2]}}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} \right]_+ = \left[\frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} \right]_+. \quad (\text{D.3.11})$$

where we used that $\varepsilon = \varepsilon_0\rho$. And, since $\frac{\varepsilon_0\rho^2}{\sigma^2} = \frac{\varepsilon_0}{\sigma_0^2} \leq \frac{\lambda_0^*}{8}$ by Assumption D.4, λ^* actually satisfies

$$\frac{\lambda_0^*}{8\rho} \leq \lambda^* \leq \frac{\lambda_0^*}{4\rho}. \quad (\text{D.3.12})$$

Moreover, Lemma D.7.8 also shows that, on $[0, 2\lambda^*]$, $\overline{\psi}_\rho$ is strongly convex with modulus

$$\frac{\mathbb{E}_P[\|\nabla f\|_2^2]}{(2\lambda^* + \frac{\varepsilon_0\rho}{\sigma^2})^3} + \frac{\varepsilon_0 d\rho}{2(2\lambda^* + \frac{\varepsilon_0\rho}{\sigma^2})^2} = \frac{\mathbb{E}_P[\|\nabla f\|_2^2]}{(\frac{\lambda_0^*}{2\rho} - \frac{\varepsilon_0\rho}{\sigma^2})^3} + \frac{\varepsilon_0 d\rho}{2(\frac{\lambda_0^*}{2\rho} - \frac{\varepsilon_0\rho}{\sigma^2})^2} \geq \rho^3 \mu^*. \quad (\text{D.3.13})$$

Now, we notice that $\varepsilon = \varepsilon_0\rho \leq \varepsilon_1$ by Assumption D.4. Then, if $\lambda \in [\lambda_1 + L_2, 2\lambda^* - L_2]$, then Lemma D.1.3 (applied twice) and the strong convexity of $\overline{\psi}_\rho$ yield

$$\lambda\rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.3.14})$$

$$\geq \lambda\rho^2 + \mathbb{E}_{\xi \sim P} [\overline{\phi}(f, \xi, \lambda + L_2, \varepsilon, \sigma)] - \varepsilon c_1 e^{-c_2 \left(\frac{\lambda + L_2}{\varepsilon} \right)^{\frac{1}{3}}} \quad (\text{D.3.15})$$

$$\geq \lambda^* \rho^2 + \mathbb{E}_{\xi \sim P} [\overline{\phi}(f, \xi, \lambda^*, \varepsilon, \sigma)] - \rho^2 L_2 + \frac{\rho^3 \mu^*}{2} (\lambda^* - (\lambda + L_2))^2 - \varepsilon c_1 e^{-c_2 \left(\frac{\lambda + L_2}{\varepsilon} \right)^{\frac{1}{3}}} \quad (\text{D.3.16})$$

$$\geq \lambda^* \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda^* + L_2, \varepsilon, \sigma)] - \rho^2 L_2 + \frac{\rho^3 \mu^*}{2} (\lambda^* - (\lambda + L_2))^2 - \varepsilon c_1 e^{-c_2 \left(\frac{\lambda + L_2}{\varepsilon} \right)^{\frac{1}{3}}} - \varepsilon c_1 e^{-c_2 \left(\frac{\lambda^*}{\varepsilon} \right)^{\frac{1}{3}}}. \quad (\text{D.3.17})$$

We first wish to choose $\lambda = \frac{\lambda^*}{2} - L_2$. By (D.3.12), since $\rho \leq \frac{\lambda_0^*}{32(\lambda_1 + L_2)}$ by Assumption D.4, this choice of λ is indeed greater than or equal to $\lambda_1 + L_2$ and (D.3.17) leads to

$$\left(\frac{\lambda^*}{2} - L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim P} \left[\phi(f, \xi, \frac{\lambda^*}{2} - L_2, \varepsilon, \sigma) \right] \quad (\text{D.3.18})$$

$$\geq \lambda^* \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda^* + L_2, \varepsilon, \sigma)] - \rho^2 L_2 + \frac{\rho^3 \mu^*}{8} (\lambda^*)^2 - \varepsilon c_1 e^{-c_2 \left(\frac{\lambda^*}{2\varepsilon} \right)^{\frac{1}{3}}} - \varepsilon c_1 e^{-c_2 \left(\frac{\lambda^*}{\varepsilon} \right)^{\frac{1}{3}}} \quad (\text{D.3.19})$$

$$\geq (\lambda^* + L_2) \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda^* + L_2, \varepsilon, \sigma)] - 2\rho^2 L_2 + \frac{\rho \mu^*}{512} (\lambda_0^*)^2 - 2\varepsilon c_1 e^{-c_2 \left(\frac{\lambda_0^*}{8\varepsilon\rho} \right)^{\frac{1}{3}}}. \quad (\text{D.3.20})$$

where we used (D.3.12) again for the last inequality.

To obtain the other inequality we pick $\lambda = 2\lambda^* - L_2$, which is greater or equal to $\lambda_1 + L_2$ by Assumption D.4 and (D.3.12) as above. Then, (D.3.17) yields

$$(2\lambda^* - L_2) \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, 2\lambda^* - L_2, \varepsilon, \sigma)] \quad (\text{D.3.21})$$

$$\geq \lambda^* \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda^* + L_2, \varepsilon, \sigma)] - \rho^2 L_2 + \frac{\rho^3 \mu^*}{2} (\lambda^*)^2 - \varepsilon c_1 e^{-c_2 \left(\frac{2\lambda^*}{\varepsilon}\right)^{\frac{1}{3}}} - \varepsilon c_1 e^{-c_2 \left(\frac{\lambda^*}{\varepsilon}\right)^{\frac{1}{3}}} \quad (\text{D.3.22})$$

$$\geq (\lambda^* + L_2) \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda^* + L_2, \varepsilon, \sigma)] - 2\rho^2 L_2 + \frac{\rho \mu^*}{512} (\lambda_0^*)^2 - 2\varepsilon c_1 e^{-c_2 \left(\frac{\lambda_0^*}{8\varepsilon\rho}\right)^{\frac{1}{3}}} \quad (\text{D.3.23})$$

where we used (D.3.12) again, and degraded the constants to match those of (D.3.20).

Thus, we have that $\lambda^* = \frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2}$ and

$$(\lambda^* + L_2) \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda^* + L_2, \varepsilon, \sigma)] - 2\rho^2 L_2 + \frac{\rho \mu^*}{512} (\lambda_0^*)^2 - 2\varepsilon c_1 e^{-c_2 \left(\frac{\lambda_0^*}{8\varepsilon\rho}\right)^{\frac{1}{3}}} \quad (\text{D.3.24})$$

$$\leq \min \left(\left(\frac{\lambda^*}{2} - L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim P} \left[\phi(f, \xi, \frac{\lambda^*}{2} - L_2, \varepsilon, \sigma) \right], (2\lambda^* - L_2) \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, 2\lambda^* - L_2, \varepsilon, \sigma)] \right) \quad (\text{D.3.25})$$

All that is left to show for the main result of the lemma is that

$$-2\rho^2 L_2 + \frac{\rho \mu^*}{512} (\lambda_0^*)^2 - 2\varepsilon c_1 e^{-c_2 \left(\frac{\lambda_0^*}{8\varepsilon\rho}\right)^{\frac{1}{3}}} \geq \frac{\rho \mu^*}{1024} (\lambda_0^*)^2 \quad (\text{D.3.26})$$

$$\Leftrightarrow 2\rho L_2 + 2\varepsilon_0 c_1 e^{-c_2 \left(\frac{\lambda_0^*}{8\varepsilon_0\rho^2}\right)^{\frac{1}{3}}} \leq \frac{\mu^*}{1024} (\lambda_0^*)^2. \quad (\text{D.3.27})$$

This is a consequence of Assumption D.4 which states that

$$\rho \leq \frac{\mu^*}{4096L_2} (\lambda_0^*)^2 \quad \text{and} \quad \rho \leq \sqrt{\frac{c_2^3 \lambda_0^*}{8\varepsilon_0}} \left(\log \left(\frac{4096\varepsilon_0 c_1}{\mu^* (\lambda_0^*)^2} \right) \right)_+^{-\frac{3}{2}}, \quad (\text{D.3.28})$$

$$\text{which imply that } 2\rho L_2 \leq \frac{\mu^*}{2048} (\lambda_0^*)^2, \quad \text{and} \quad 2\varepsilon_0 c_1 e^{-c_2 \left(\frac{\lambda_0^*}{8\varepsilon_0\rho^2}\right)^{\frac{1}{3}}} \leq \frac{\mu^*}{2048} (\lambda_0^*)^2, \quad (\text{D.3.29})$$

so that (D.3.27) indeed holds, concluding the proof of the first part of the result.

The supplementary bounds follow directly from (D.3.12) and our assumptions on ρ . $\blacksquare$

We are now in a position to show our main result when ρ is small, namely Proposition D.3.1.

Proof of Proposition D.3.1. Let us first take any $\lambda \in [\max(\frac{\lambda_0^*}{32\rho}, \lambda_1 + L_2), \frac{\lambda_0^*}{2\rho}]$. We want to instantiate Lemma D.7.2 with $X(f, \xi) \leftarrow \phi(f, \xi, \lambda, \varepsilon, \sigma)$, $(\mathcal{X}, \text{dist}) \leftarrow (\mathcal{F}, \|\cdot\|_\infty)$, whose requirements are checked since:

1. For any $f \in \mathcal{F}$, $\phi(f, \xi, \lambda, \varepsilon, \sigma)$ is measurable since the functions of $\mathcal{F}$ are continuous and thus *a fortiori* measurable;
2. By the proof of [Lemma D.2.2\(a\)](#), we have that for any $\varepsilon \geq 0$, $\sigma > 0$ and any $\xi \in \text{supp } P$, $f \mapsto \phi(f, \xi, \lambda, \varepsilon, \sigma)$ is 1-Lipschitz continuous w.r.t. the norm $\|\cdot\|_\infty$;
3. With [Proposition D.1.2](#) with $\underline{\lambda} \leftarrow \max\left(\frac{\lambda_0^*}{32\rho}, \lambda_1 + L_2\right)$, $\bar{\lambda} \leftarrow \frac{\lambda_0^*}{2\rho}$, for any $f \in \mathcal{F}$, $\xi \in \text{supp } P$ and $\varepsilon \in [0, \varepsilon_1]$ (by [Assumption D.4](#)), we have

$$-\bar{M}(\rho) \leq f(\xi) - M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \leq \phi(f, \xi, \lambda, \varepsilon, \sigma) \leq f(\xi) + M(\underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \leq \tilde{F}(0) + \bar{M}(\rho) \quad (\text{D.3.30})$$

where $\bar{M}(\rho)$ is defined in [Proposition D.3.1](#).

Since $\frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2 \geq \lambda_1 + L_2$ by [Lemma D.3.2](#), we can apply statement (b) of [Lemma D.7.2](#) with $\lambda \leftarrow \frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2$ and $\delta \leftarrow \frac{\delta}{4}$ to have that, with probability at least $1 - \frac{\delta}{4}$, for all $f \in \mathcal{F}$

$$\mathbb{E}_{\xi \sim P} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2, \varepsilon \right) \right] - \mathbb{E}_{\xi \sim \hat{P}_n} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2, \varepsilon \right) \right] \quad (\text{D.3.31})$$

$$\leq \frac{48\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\sqrt{n}} + 4(\tilde{F}(0) + \bar{M}(\rho)) \sqrt{\frac{\log \frac{4}{\delta}}{2n}}. \quad (\text{D.3.32})$$

Similarly, we can apply statement (a) of [Lemma D.7.2](#) with $\lambda \leftarrow \frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2$ and $\delta \leftarrow \frac{\delta}{4}$ to get that, with probability at least $1 - \frac{\delta}{4}$, for all $f \in \mathcal{F}$,

$$\mathbb{E}_{\xi \sim \hat{P}_n} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2, \varepsilon \right) \right] - \mathbb{E}_{\xi \sim P} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2, \varepsilon \right) \right] \quad (\text{D.3.33})$$

$$\leq \frac{48\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\sqrt{n}} + 4(\tilde{F}(0) + \bar{M}(\rho)) \sqrt{\frac{\log \frac{4}{\delta}}{2n}}. \quad (\text{D.3.34})$$

Combining the two statements above and using [Lemma D.3.2](#), we get that, with probability at least $1 - \frac{\delta}{2}$, for any $f \in \mathcal{F}$,

$$\left(\frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim \hat{P}_n} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2, \varepsilon \right) \right] \quad (\text{D.3.35})$$

$$\geq \left(\frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim P} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2, \varepsilon \right) \right] \quad (\text{D.3.36})$$

$$- \frac{48\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\sqrt{n}} - 4(\tilde{F}(0) + \bar{M}(\rho)) \sqrt{\frac{\log \frac{4}{\delta}}{2n}} \quad (\text{D.3.37})$$

$$\geq \left(\frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim P} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2, \varepsilon \right) \right] \quad (\text{D.3.38})$$

$$+ \frac{\rho\mu^*}{1024} (\lambda_0^*)^2 - \frac{48\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\sqrt{n}} - 4(\tilde{F}(0) + \bar{M}(\rho)) \sqrt{\frac{\log \frac{4}{\delta}}{2n}} \quad (\text{D.3.39})$$

$$\geq \left(\frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2 \right) \rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} \left[\phi \left(f, \xi, \frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2, \varepsilon \right) \right] \quad (\text{D.3.40})$$

$$+ \frac{\rho\mu^*}{1024} (\lambda_0^*)^2 - \frac{96\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\sqrt{n}} - 8(\widetilde{F}(0) + \overline{M}(\rho)) \sqrt{\frac{\log \frac{4}{\delta}}{2n}}. \quad (\text{D.3.41})$$

Noting that the assumption on ρ in [Proposition D.3.1](#) implies that

$$\frac{\rho\mu^*}{1024} (\lambda_0^*)^2 \geq \frac{96\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\sqrt{n}} + 8 \left(\widetilde{F}(0) + \overline{M}(\rho) \right) \sqrt{\frac{\log \frac{4}{\delta}}{2n}} \quad (\text{D.3.42})$$

we have proven that, with probability at least $1 - \frac{\delta}{2}$, for any $f \in \mathcal{F}$,

$$\psi_\rho \left(\frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2 \right) \geq \psi_\rho \left(\frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2 \right) \quad (\text{D.3.43})$$

where $\psi_\rho : \lambda \mapsto \lambda\rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]$. Now, since ψ_ρ is convex, this means that its minimizers on $\mathbb{R}_+$ are greater than

$$\frac{\lambda_0^*}{8\rho} - \frac{\varepsilon_0\rho}{2\sigma^2} - L_2 \geq \frac{\lambda_0^*}{32\rho} \quad (\text{D.3.44})$$

where the inequality comes from [Lemma D.3.2](#).

Using the same reasoning, one can get that with probability at least $1 - \frac{\delta}{2}$ the minimizers are no greater than

$$\frac{\lambda_0^*}{4\rho} - \frac{\varepsilon_0\rho}{\sigma^2} + L_2 \leq \frac{\lambda_0^*}{2\rho}. \quad (\text{D.3.45})$$

Thus, we have shown that with probability at least $1 - \frac{\delta}{2}$, for any $f \in \mathcal{F}$,

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\lambda \geq 0} \lambda\rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.3.46})$$

$$= \inf_{\frac{\lambda_0^*}{32\rho} \leq \lambda \leq \frac{\lambda_0^*}{2\rho}} \lambda\rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] . \quad (\text{D.3.47})$$

■

D.4 DUAL BOUND WHEN ρ IS CLOSE TO THIS MAXIMAL RADIUS

Complementary to the previous section [Appendix D.3](#), we consider the case where ρ is close than the critical radius. Though the bounds of this section are much worse than the one of [Appendix D.3](#) when ρ goes to zero, they hold for the whole ranges of ρ considered in the theorems.

As mentioned in [Remark 7](#), as ρ grows, the Wasserstein ball constraint can stop being active, leading to a null dual variable. Thus, it is essential that ρ be lower than the critical radius to stay in the distributionally robust regime and to avoid the worst-case regime. In that case, we are able to lower-bound the dual multiplier λ .

We defined the critical radius in standard WDRO case in [Remark 7](#) and we extend it here to cover the regularized case:

$$\rho_c^2(\varepsilon, \sigma) := \begin{cases} \inf_{f \in \mathcal{F}} \mathbb{E}_{\xi \sim P} \left[\mathbb{E}_{\zeta \sim \pi_\sigma^{f/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \right] & \text{if } \varepsilon > 0 \\ \inf_{f \in \mathcal{F}} \mathbb{E}_{\xi \sim P} \left[\min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max f \right\} \right] & \text{otherwise} \end{cases} \quad (\text{D.4.1})$$

where $\pi_\sigma^g(\mathrm{d}\zeta|\xi) \propto e^{g(\zeta)} \pi_\sigma(\mathrm{d}\zeta|\xi)$ is a conditional probability distribution parametrized by an $\Xi \rightarrow \mathbb{R}$ function g , i.e.,

$$\mathbb{E}_{\zeta \sim \pi_\sigma^g(\cdot|\xi)} [h(\xi, \zeta)] = \frac{\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} [e^{g(\zeta)} h(\xi, \zeta)]}{\mathbb{E}_{\zeta' \sim \pi_\sigma(\cdot|\xi)} [e^{g(\zeta')}]} \quad (\text{D.4.2})$$

For this part of the proof, the case when $\varepsilon = 0$ differs from the regularized one $\varepsilon > 0$. We thus present them in separate sections [Appendices D.4.1](#) and [D.4.2](#).

D.4.1 Standard WDRO case

The main result of this section in the standard WDRO case is [Proposition D.4.1](#) below.

Proposition D.4.1. *Let [Assumption 6.5](#) hold and fix a threshold $\delta \in (0, 1)$. Assume that*

$$\rho^2 \leq \rho_c^2(0, 0) - \frac{2B(\delta)}{\sqrt{n}} \quad (\text{D.4.3})$$

$$\text{with } L := \frac{16 \sup_{f \in \mathcal{F}} \mathbb{E}_{\xi \sim P} \left[\frac{1}{2} d^2(\xi, \arg \max f) \right]}{\mu} \quad (\text{D.4.4})$$

$$\text{and } B(\delta) := 48\mathcal{I}(\mathcal{F}, D) + 2\sqrt{C^* \log 1/\delta}. \quad (\text{D.4.5})$$

where $\mathcal{I}(\mathcal{F}, D)$ and D are defined in [Appendix D.1.4](#), and $\lambda_2 > 0$ is a constant depending on $\Xi, \mathcal{F}, L_3, \mu$ and C^* .

Then, with probability at least $1 - \delta$, we have

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda} \leq \lambda} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, 0)] \quad (\text{D.4.6})$$

where the dual bound $\underline{\lambda}$ is defined as

$$\underline{\lambda} := \min \left(\lambda_2, \frac{\rho_c^2(0, 0) - \rho^2}{2L} \right). \quad (\text{D.4.7})$$

Before proceeding with the proof, we need to prove the following lemma which leverages [Assumption 6.5](#).

Lemma D.4.2. *Fix $f \in \mathcal{F}$ and $\xi \in \text{supp } P$. There exists a constant $\lambda_2 > 0$ depending on $\Xi, \mathcal{F}, L_3, \mu$ and C^* such that, for $\lambda \in [0, \lambda_2]$,*

$$\min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max_{\Xi} f - \frac{\lambda}{2} \|\xi - \cdot\|^2 \right\} \geq \left(1 - \frac{16\lambda}{\mu} \right) \min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max_{\Xi} f \right\}. \quad (\text{D.4.8})$$

Proof. Fix $f \in \mathcal{F}$ and $\xi \in \text{supp } P$. Define, for convenience, $\Xi^* := \arg \max f$ and

$$Y(\lambda) := \min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max_{\Xi} f - \frac{\lambda}{2} \|\xi - \cdot\|^2 \right\}. \quad (\text{D.4.9})$$

Step 1: Localization in a $\mathcal{O}(1)$ -neighborhood of Ξ^* . For a fixed $R^* > 0$ that will be chosen later, we show that, for λ small enough, $Y(\lambda)$ is equal to

$$\min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max_{(\Xi^*)^{R^*}} f - \frac{\lambda}{2} \|\xi - \cdot\|^2 \right\} \quad \text{where} \quad (\Xi^*)^{R^*} := \{\xi \in \Xi : d(\xi, \Xi^*) \leq R^*\}. \quad (\text{D.4.10})$$

Indeed, by [Assumption 6.5](#), there is some $\Delta(R^*) > 0$ such that for all $f \in \mathcal{F}$ and $\zeta \in \Xi \setminus (\Xi^*)^{R^*}$,

$$f(\zeta) - \max f - \frac{\lambda}{2} \|\xi - \zeta\|^2 \leq f(\zeta) - \max f \leq -\Delta(R^*), \quad (\text{D.4.11})$$

while, for any $\xi^* \in \Xi^*$,

$$f(\xi^*) - \max f - \frac{\lambda}{2} \|\xi - \xi^*\|^2 = -\frac{\lambda}{2} \|\xi - \xi^*\|^2 \geq -\lambda C^*. \quad (\text{D.4.12})$$

Hence, for $\lambda \leq \frac{\Delta(R^*)}{C^*}$, $f(\zeta) - \frac{\lambda}{2} \|\xi - \zeta\|^2 \leq -\Delta(R^*) + \max f \leq -\lambda C^* + \max f \leq f(\xi^*) - \frac{\lambda}{2} \|\xi - \xi^*\|^2$. This means that points in $\Xi \setminus (\Xi^*)^{R^*}$ cannot maximize $f - \frac{\lambda}{2} \|\xi - \cdot\|^2$ and so it suffices to consider the $\arg \max$ over $(\Xi^*)^{R^*}$ in the definition of $Y(\lambda)$.

Step 2: Localization in a $\mathcal{O}(\lambda)$ -neighborhood of Ξ^* . Take $\zeta^* \in \arg \max_{(\Xi^*)^{R^*}} f - \frac{\lambda}{2} \|\xi - \cdot\|^2$. Since $d(\zeta^*, \Xi^*) \leq R^*$, the Euclidean projection of ζ^* on Ξ^* , that we denote by ξ^* , is at most at distance R^* of ζ^* and $\zeta^* - \xi^* \in \widehat{\text{NC}}_{\xi^*}(\Xi^*)$, see e.g., [Rockafellar and Wets \(1998, Thm. 6.12\)](#). By the growth condition of [Assumption 6.5](#), we get that

$$f(\xi^*) \geq f(\zeta^*) + \frac{\mu}{2} \|\zeta^* - \xi^*\|^2 - \frac{L_3}{6} \|\zeta^* - \xi^*\|^3. \quad (\text{D.4.13})$$

But, by definition of ζ^* , we also have that

$$f(\zeta^*) - \frac{\lambda}{2} \|\xi - \zeta^*\|^2 \geq f(\xi^*) - \frac{\lambda}{2} \|\xi - \xi^*\|^2. \quad (\text{D.4.14})$$

Plugging [\(D.4.13\)](#) we get that

$$-\frac{\lambda}{2} \|\xi - \zeta^*\|^2 \geq -\frac{\lambda}{2} \|\xi - \xi^*\|^2 + \frac{\mu}{2} \|\zeta^* - \xi^*\|^2 - \frac{L_3}{6} \|\zeta^* - \xi^*\|^3. \quad (\text{D.4.15})$$

Rearranging and developing $\frac{1}{2} \|\xi - \xi^*\|^2$ yields

$$\frac{L_3}{6} \|\zeta^* - \xi^*\|^3 + \lambda \langle \xi - \zeta^*, \zeta^* - \xi^* \rangle \geq \frac{\lambda + \mu}{2} \|\xi^* - \zeta^*\|^2, \quad (\text{D.4.16})$$

which gives, by Cauchy-Schwarz inequality,

$$\frac{L_3}{6} \|\zeta^* - \xi^*\|^3 + \lambda \|\xi - \zeta^*\| \|\zeta^* - \xi^*\| \geq \frac{\lambda + \mu}{2} \|\xi^* - \zeta^*\|^2, \quad (\text{D.4.17})$$

We now wish to obtain a bound on $u^\star := \|\zeta^\star - \xi^\star\|$. If it is zero, there is nothing to do. Otherwise, assuming that it is positive, (D.4.17) gives the inequation

$$\frac{\mu + \lambda}{2} u^\star \leq \frac{L_3}{6} (u^\star)^2 + \lambda \|\xi - \xi^\star\|. \quad (\text{D.4.18})$$

When $\frac{(\mu + \lambda)^2}{4} - \frac{2L_3\lambda}{3} \|\xi - \xi^\star\|$ is non-negative, this inequation is satisfied for

$$u^\star \notin \left[\frac{(\mu + \lambda) \pm \sqrt{(\mu + \lambda)^2 - 8L_3\lambda \|\xi - \xi^\star\|/3}}{L_3/3} \right]. \quad (\text{D.4.19})$$

Hence, in particular, if $u^\star \leq \frac{3\mu}{L_3}$, then $u^\star$ must be less or equal than

$$\frac{(\mu + \lambda) - \sqrt{(\mu + \lambda)^2 - 8L_3\lambda \|\xi - \xi^\star\|/3}}{L_3/3} = \frac{3(\mu + \lambda)}{L_3} \left(1 - \sqrt{1 - \frac{8L_3\lambda \|\xi - \xi^\star\|}{3(\mu + \lambda)^2}} \right) \leq \frac{8\lambda \|\xi - \xi^\star\|}{\mu + \lambda} \quad (\text{D.4.20})$$

when $\frac{8L_3\lambda \|\xi - \xi^\star\|}{3(\mu + \lambda)^2} \leq 1$, using that $1 - \sqrt{1 - x} \leq x$ for $x \in [0, 1]$.

Thus, assuming that λ is small enough so that $8L_3\lambda C^\star \leq 3(\mu)^2$ and choosing $R^\star := \frac{3\mu}{L_3}$ so that $u^\star \leq \frac{3\mu}{L_3}$ by construction, we have that for any $\zeta^\star \in \arg \max_{(\Xi^\star)^{R^\star}} f - \frac{\lambda}{2} \|\xi - \cdot\|^2$, there is a point $\xi^\star \in \Xi^\star$ such that

$$\|\zeta^\star - \xi^\star\| \leq \frac{8\lambda \|\xi - \xi^\star\|}{\mu}. \quad (\text{D.4.21})$$

Step 3: Conclusion. Defining the constant

$$\lambda_2 := \min \left(\frac{\Delta(R^\star)}{C^\star}, \frac{3\mu^2}{8L_3C^\star}, \frac{\mu}{16} \right), \quad (\text{D.4.22})$$

and using the previous steps, we have for any $\lambda \in [0, \lambda_2]$ and any $\zeta^\star \in \arg \max_{\Xi} f - \frac{\lambda}{2} \|\xi - \cdot\|^2$

$$\frac{1}{2} \|\xi - \zeta^\star\|^2 = \frac{1}{2} \|\xi - \xi^\star\|^2 + \frac{1}{2} \|\xi^\star - \zeta^\star\|^2 - \langle \xi - \xi^\star, \xi^\star - \zeta^\star \rangle \quad (\text{D.4.23})$$

$$\geq \frac{1}{2} \|\xi - \xi^\star\|^2 - \|\xi - \xi^\star\| \|\xi^\star - \zeta^\star\| \quad (\text{D.4.24})$$

$$\geq \left(1 - \frac{16\lambda}{\mu} \right) \frac{1}{2} \|\xi - \xi^\star\|^2 \quad (\text{D.4.25})$$

which concludes the proof. $\blacksquare$

We can now turn to the proof of our proposition.

Proof of Proposition D.4.1. Let $0 \leq \lambda \leq \underline{\lambda}$. For $f \in \mathcal{F}$ and $\lambda \geq 0$, we define $\hat{\psi}_\rho : \lambda \mapsto \lambda \rho^2 + \mathbb{E}_{\xi \sim \hat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, 0)]$ and its (right-sided) derivative $\partial_\lambda \hat{\psi}_\rho$. This derivative is given by,

$$\partial_\lambda \hat{\psi}_\rho(\lambda) = \rho^2 - \mathbb{E}_{\xi \sim \hat{\mathbb{P}}_n} \left[\min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max_{\Xi} f - \frac{\lambda}{2} \|\xi - \cdot\|^2 \right\} \right] \quad (\text{D.4.26})$$

$$\leq \rho^2 - \left(1 - \frac{16\lambda}{\mu}\right) \mathbb{E}_{\xi \sim \hat{P}_n} \left[\min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max_{\Xi} f \right\} \right] \quad (\text{D.4.27})$$

where we used [Lemma D.4.2](#) with $\lambda \leq \underline{\lambda} \leq \lambda_2$.

We then instantiate [Lemma D.7.2](#) with $X(f, \xi) \leftarrow \frac{1}{2} d^2(\xi, \arg \max f), (\mathcal{X}, \text{dist}) \leftarrow (\mathcal{F}, D)$, whose requirements are checked since:

1. For any $f \in \mathcal{F}$, $X(f, \cdot)$ is measurable since the functions of $\mathcal{F}$ are continuous and thus $\arg \max f$ is *a fortiori* measurable;
2. By definition of D , for any $\xi \in \text{supp } P$, $f \mapsto d(\xi, \arg \max f)$ is 1-Lipschitz w.r.t. this distance so that $X(\xi, \cdot)$ is $\sqrt{2C^*}$ -Lipschitz.
3. By construction, the range of values X is included in $[0, C^*]$.

We can thus apply statement (b) of [Lemma D.7.2](#) to have that, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$\mathbb{E}_{\xi \sim \hat{P}_n} \left[\frac{1}{2} d^2(\xi, \arg \max f) \right] \geq \mathbb{E}_{\xi \sim P} \left[\frac{1}{2} d^2(\xi, \arg \max f) \right] - \frac{B(\delta)}{\sqrt{n}} \quad (\text{D.4.28})$$

Hence, putting this bound together with (D.4.27) yields

$$\partial_\lambda \hat{\psi}_\rho(\lambda) \leq \rho^2 - \left(1 - \frac{16\lambda}{\mu}\right) \mathbb{E}_{\xi \sim P} \left[\frac{1}{2} d^2(\xi, \arg \max f) \right] + \frac{B(\delta)}{\sqrt{n}} \quad (\text{D.4.29})$$

$$\leq \rho^2 - \rho_c^2 + L\lambda + \frac{B(\delta)}{\sqrt{n}}, \quad (\text{D.4.30})$$

which is non-negative for $\lambda \leq \underline{\lambda}$. ■

D.4.2 Regularized case

The main bound on λ of this section are given by [Proposition D.4.3](#).

Proposition D.4.3. Fix a threshold $\delta \in (0, 1)$. Assume that $\varepsilon_0 > 0$ and that

$$\rho^2 \leq \rho_c^2(\varepsilon, \sigma) - \left(\frac{48\sqrt{\text{Var}(\varepsilon, \sigma)}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\varepsilon\sqrt{n}} + 2C_{\mathcal{F}}(\varepsilon, \sigma)\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \right) \quad (\text{D.4.31})$$

where $\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)$ is defined in [Appendix D.1.1](#).

Then, with probability at least $1 - \delta$, we have

$$\forall f \in \mathcal{F}, \quad \hat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda}_n \leq \lambda \leq \bar{\lambda}} \lambda \rho^2 + \mathbb{E}_{\xi \sim \hat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.4.32})$$

where the dual bounds are defined by

$$\underline{\lambda}_n := \frac{\varepsilon}{\text{Var}(\varepsilon, \sigma)} \left(\rho_c^2(\varepsilon, \sigma) - \rho^2 - \left(\frac{48\sqrt{\text{Var}(\varepsilon, \sigma)}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\varepsilon\sqrt{n}} + 2C_{\mathcal{F}}(\varepsilon, \sigma)\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \right) \right) \quad (\text{D.4.33})$$

$$\text{and } \bar{\lambda} := \max \left(\frac{12\varepsilon}{R^2} \log(2 \times 6^{d/2}), e^{\frac{\sup_{f \in \mathcal{F}} \|f\|_\infty}{\varepsilon}} \frac{\varepsilon_0}{\rho} \right). \quad (\text{D.4.34})$$

Proof. Lower-bound: By [Assumption D.1](#), for any $f \in \mathcal{F}$, $\xi \in \text{supp } P$, $\lambda \mapsto \phi(f, \xi, \lambda, \varepsilon, \sigma)$ is twice differentiable and its derivatives are for any $\lambda \geq 0$

$$\partial_\lambda \phi(f, \xi, \lambda, \varepsilon, \sigma) = -\mathbb{E}_{\zeta \sim \pi_\sigma^{\frac{f-\lambda\|\xi-\zeta\|^2/2}{\varepsilon}}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \quad (\text{D.4.35})$$

$$\partial_\lambda^2 \phi(f, \xi, \lambda, \varepsilon, \sigma) = \frac{1}{\varepsilon} \text{Var}_{\zeta \sim \pi_\sigma^{\frac{f-\lambda\|\xi-\zeta\|^2/2}{\varepsilon}}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right], \quad (\text{D.4.36})$$

and using $\text{Var}(\varepsilon, \sigma)$ which is defined in [Appendix D.1.1](#), we get that, for any $\lambda \geq 0$,

$$0 \leq \partial_\lambda^2 \phi(f, \xi, \lambda, \varepsilon, \sigma) \leq \frac{1}{\varepsilon} \text{Var}(\varepsilon, \sigma). \quad (\text{D.4.37})$$

As a consequence,

$$\partial_\lambda \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \right\} = \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\partial_\lambda \phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.4.38})$$

$$\leq \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\partial_\lambda \phi(f, \xi, 0, \varepsilon, \sigma)] + \frac{\lambda}{\varepsilon} \text{Var}(\varepsilon, \sigma) \quad (\text{D.4.39})$$

$$= \rho^2 - \mathbb{E}_{\xi \sim \widehat{P}_n} \left[\mathbb{E}_{\zeta \sim \pi_\sigma^{f/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \right] + \frac{\lambda}{\varepsilon} \text{Var}(\varepsilon, \sigma). \quad (\text{D.4.40})$$

Now, we want to instantiate [Lemma D.7.2](#) with $X(f, \xi) := \mathbb{E}_{\zeta \sim \pi_\sigma^{f/\varepsilon}(\cdot|\xi)} \frac{1}{2} \|\xi - \zeta\|^2$, $(\mathcal{X}, \text{dist}) \leftarrow (\mathcal{F}, \|\cdot\|_\infty)$, whose requirements are checked since:

1. For any $f \in \mathcal{F}$, $\mathbb{E}_{\zeta \sim \pi_\sigma^{f/\varepsilon}(\cdot|\xi)} \frac{1}{2} \|\xi - \zeta\|^2$ is measurable since the functions of $\mathcal{F}$ are continuous and thus *a fortiori* measurable;
2. To show that $f \mapsto X(f, \xi)$ is $\frac{1}{\varepsilon} \sqrt{\text{Var}(\varepsilon, \sigma)}$ -Lipschitz, we take $f, g \in \mathcal{F}$ and define, for $t \in [0, 1]$, $f_t = f + t(g - f)$. Since, $\|f - g\|_\infty < +\infty$ and $\sup_{\xi \in \text{supp } P} \mathbb{E}_{\zeta \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} \frac{1}{2} \|\xi - \zeta\|^2 < +\infty$ by compactness of Ξ , [Assumption 6.1](#), $t \mapsto X(f_t, \xi)$ is differentiable with derivative,

$$\frac{d}{dt} X(f_t, \xi) = \frac{1}{\varepsilon} \mathbb{E}_{\zeta \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 (g(\zeta) - f(\zeta)) \right] \quad (\text{D.4.41})$$

$$- \frac{1}{\varepsilon} \mathbb{E}_{\zeta \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \mathbb{E}_{\zeta \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} [g(\zeta) - f(\zeta)] \quad (\text{D.4.42})$$

$$= \frac{1}{\varepsilon} \mathbb{E}_{\zeta \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} \left[\left(\frac{1}{2} \|\xi - \zeta\|^2 - \mathbb{E}_{\zeta' \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta'\|^2 \right] \right) \right] \quad (\text{D.4.43})$$

$$\times \left((g(\zeta) - f(\zeta)) - \mathbb{E}_{\zeta' \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} [g(\zeta') - f(\zeta')] \right). \quad (\text{D.4.44})$$

By using Cauchy-Schwarz inequality, we get that,

$$\frac{d}{dt} X(f_t, \xi) \leq \frac{1}{\varepsilon} \sqrt{\text{Var}_{\zeta \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right]} \sqrt{\text{Var}_{\zeta \sim \pi_\sigma^{f_t/\varepsilon}(\cdot|\xi)} [g(\zeta) - f(\zeta)]} \quad (\text{D.4.45})$$

$$\leq \frac{1}{\varepsilon} \sqrt{\text{Var}_{\zeta \sim \pi_{\sigma}^{f/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right]} \sqrt{\mathbb{E}_{\zeta \sim \pi_{\sigma}^{f/\varepsilon}(\cdot|\xi)} \left[(g(\zeta) - f(\zeta))^2 \right]} \quad (\text{D.4.46})$$

$$\leq \frac{1}{\varepsilon} \sqrt{\text{Var}_{\zeta \sim \pi_{\sigma}^{f/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right]} \|g - f\|_{\infty}, \quad (\text{D.4.47})$$

which gives the desired Lipschitz condition;

3. The random variables $X(f, \xi)$ lie between 0 and $C_{\mathcal{F}}(\varepsilon, \sigma)$, which is defined in [Appendix D.1.1](#).

We can thus apply statement (b) of [Lemma D.7.2](#) to have that, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$

$$\mathbb{E}_{\xi \sim P} \left[\mathbb{E}_{\zeta \sim \pi_{\sigma}^{f/\varepsilon}(\cdot|\xi)} \frac{1}{2} \|\xi - \zeta\|^2 \right] - \mathbb{E}_{\xi \sim \hat{P}_n} \left[\mathbb{E}_{\zeta \sim \pi_{\sigma}^{f/\varepsilon}(\cdot|\xi)} \frac{1}{2} \|\xi - \zeta\|^2 \right] \quad (\text{D.4.48})$$

$$\leq \frac{48\sqrt{\text{Var}(\varepsilon, \sigma)}\mathcal{I}(\mathcal{F}, \|\cdot\|_{\infty})}{\varepsilon\sqrt{n}} + 2C_{\mathcal{F}}(\varepsilon, \sigma)\sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \quad (\text{D.4.49})$$

Combining [\(D.4.40\)](#) and [\(D.4.49\)](#), we obtain that with probability at least $1 - \delta$

$$\partial_{\lambda} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \hat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \right\} \quad (\text{D.4.50})$$

$$\leq \rho^2 - \mathbb{E}_{\xi \sim P} \left[\mathbb{E}_{\zeta \sim \pi_{\sigma}^{f/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \right] + \frac{\lambda}{\varepsilon} \text{Var}(\varepsilon, \sigma) + \frac{48\sqrt{\text{Var}(\varepsilon, \sigma)}\mathcal{I}(\mathcal{F}, \|\cdot\|_{\infty})}{\varepsilon\sqrt{n}} + 2C_{\mathcal{F}}(\varepsilon, \sigma)\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \quad (\text{D.4.51})$$

$$\leq \rho^2 - \rho_c^2(\varepsilon, \sigma) + \frac{\lambda}{\varepsilon} \text{Var}(\varepsilon, \sigma) + \frac{48\sqrt{\text{Var}(\varepsilon, \sigma)}\mathcal{I}(\mathcal{F}, \|\cdot\|_{\infty})}{\varepsilon\sqrt{n}} + 2C_{\mathcal{F}}(\varepsilon, \sigma)\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \quad (\text{D.4.52})$$

$$= \frac{1}{\varepsilon} \text{Var}(\varepsilon, \sigma) (\lambda - \underline{\lambda}_n) \quad (\text{D.4.53})$$

where $\underline{\lambda}_n \geq 0$ is as defined in the statement of the result.

Hence, for all $0 \leq \lambda \leq \underline{\lambda}_n$, the derivative of $\lambda \mapsto \lambda \rho^2 + \mathbb{E}_{\xi \sim \hat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]$ is negative; and since this function is convex, this means that its minimizers are greater than $\underline{\lambda}_n$ with probability at least $1 - \delta$ which is our result.

Upper-bound: Almost surely, for any $f \in \mathcal{F}$, let us begin by bounding the $\partial_{\lambda} \phi(f, \xi, \lambda, \varepsilon, \sigma)$ for $\lambda \geq 0$, $f \in \mathcal{F}$ and $\xi \in \text{supp } P$. Its expression is given by

$$-\partial_{\lambda} \phi(f, \xi, \lambda, \varepsilon, \sigma) = \mathbb{E}_{\zeta \sim \pi_{\sigma}^{\frac{f - \lambda \|\xi - \cdot\|^2/2}{\varepsilon}}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \leq e^{\frac{\|f\|_{\infty}}{\varepsilon}} \frac{\int_{\Xi} \frac{1}{2} \|\xi - \zeta\|^2 e^{-\left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2}\right) \frac{1}{2} \|\xi - \zeta\|^2} d\zeta}{\int_{\Xi} e^{-\left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2}\right) \frac{1}{2} \|\xi - \zeta\|^2} d\zeta}. \quad (\text{D.4.54})$$

On the one hand, we lower-bound the denominator using [Lemma D.7.1](#) and [Assumption 6.3](#) as

$$\frac{1}{(2\pi)^{d/2}} \left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2} \right)^{d/2} \int_{\Xi} e^{-\left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2}\right) \frac{1}{2} \|\xi - \zeta\|^2} d\zeta \geq 1 - 6^{d/2} e^{-\frac{R^2}{12} \left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2} \right)} \geq \frac{1}{2}, \quad (\text{D.4.55})$$

where we used that $\lambda \geq \frac{12\varepsilon}{R^2} \log(2 \times 6^{d/2})$.

On the other hand, the denominator is upper-bounded as

$$\frac{1}{(2\pi)^{d/2}} \left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2} \right)^{d/2} \int_{\Xi} \frac{1}{2} \|\xi - \zeta\|^2 e^{-\left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2}\right) \frac{1}{2} \|\xi - \zeta\|^2} d\zeta \leq \frac{1}{2} \left(\frac{\lambda}{\varepsilon} + \frac{1}{\sigma^2} \right)^{-1} \leq \frac{\varepsilon}{2\lambda}. \quad (\text{D.4.56})$$

Hence, we have shown that $-\partial_\lambda \phi(f, \xi, \lambda, \varepsilon, \sigma) \leq e^{\frac{\|f\|_\infty}{\varepsilon}} \frac{\varepsilon}{\lambda}$ and, as a consequence,

$$\rho^2 + \mathbb{E}_{\xi \sim \hat{\mathbb{P}}_n} [\partial_\lambda \phi(f, \xi, \lambda, \varepsilon, \sigma)] \geq \rho^2 - e^{\frac{\|f\|_\infty}{\varepsilon}} \frac{\varepsilon}{\lambda}, \quad (\text{D.4.57})$$

which is non-negative for $\lambda \geq e^{\frac{\|f\|_\infty}{\varepsilon}} \frac{\varepsilon_0}{\rho}$.

Hence, for

$$\lambda \geq \bar{\lambda} := \max \left(\frac{12\varepsilon}{R^2} \log(2 \times 6^{d/2}), e^{\frac{\sup_{f \in \mathcal{F}} \|f\|_\infty}{\varepsilon}} \frac{\varepsilon_0}{\rho} \right), \quad (\text{D.4.58})$$

the derivative of $\lambda \mapsto \lambda \rho^2 + \mathbb{E}_{\xi \sim \hat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]$ is non-negative, which means that its minimizers are smaller than $\bar{\lambda}$. ■

D.5 PROOF OF THE MAIN RESULTS

In this section, we present our main results with explicit constants. In [Appendix D.5.1](#) we treat the case of standard WDRO, i.e., the setting of [Theorems 6.1](#) and [6.2](#), while in [Appendix D.5.2](#) we handle the regularized setting of [Theorem 6.6](#).

D.5.1 Standard WDRO case

The main results of this section are [Theorems D.1](#) and [D.3](#) which are more precise versions of [Theorems 6.1](#) and [6.2](#) respectively.

Theorem D.1 (Extended version of [Theorem 6.1](#)). *Under [Assumptions 6.1](#), [6.3](#) and [D.1](#) and the additional [Assumptions 6.4](#) and [6.5](#), with $\rho_c = \rho_c(0, 0)$ defined in [\(D.4.1\)](#) for any $\delta \in (0, 1)$ and $n \geq 1$, if*

$$\max \left(\rho_n, \frac{8192}{\sqrt{n} \mu^*(\lambda_0^*)^2} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \left(\tilde{F}(0) + \overline{M}(\rho_c) \right) \sqrt{1 + \log \frac{4}{\delta}} \right) \right) \leq \rho \quad (\text{D.5.1})$$

$$\text{and} \quad \rho \leq \frac{\rho_c}{2} - \frac{96\mathcal{I}(\mathcal{F}, D) + 4\sqrt{C^* \log 1/\delta}}{\sqrt{n}}. \quad (\text{D.5.2})$$

where

$$\rho_{\text{thres}} := \min \left(\frac{\lambda_0^*}{8(\lambda_1 + L_2)}, \frac{\mu^*(\lambda_0^*)^2}{4096L_2} \right) \quad (\text{D.5.3})$$

$$\bar{a} := \sup_{0 < \rho' \leq \rho_{\text{thres}}} a \left(\frac{1}{\rho'} \min \left(\frac{\lambda_0^*}{32}, \rho_{\text{thres}} \lambda_2, \frac{3\rho_c^2 \rho_{\text{thres}}}{8L} \right), \frac{\lambda_0^*}{2\rho'}, 0, 0 \right) \quad (\text{D.5.4})$$

$$\rho_n := \frac{117 \left(\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \max \left(\tilde{F} \left(\frac{\lambda_0^*}{32\rho_c} \right), \bar{a} \right) \left(1 + \sqrt{\log \frac{1}{\delta}} \right) \right)}{\sqrt{n} \min \left(\frac{\lambda_0^*}{32}, \rho_{\text{thres}} \lambda_2, \frac{3\rho_c^2 \rho_{\text{thres}}}{8L} \right)}, \quad (\text{D.5.5})$$

then, with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathbb{E}_{\xi \sim Q} [f(\xi)] \quad \text{for all } Q \text{ such that } W_2^2(P, Q) \leq \rho(\rho - \rho_n). \quad (\text{D.5.6})$$

In particular, with probability $1 - \delta$, we have

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathbb{E}_{\xi \sim P} [f(\xi)]. \quad (\text{D.5.7})$$

The proof of [Theorem D.1](#) relies on [Lemma D.5.2](#) that combines the results of the previous sections, namely propositions [D.2.1](#), [D.3.1](#) and [D.4.1](#).

Lemma D.5.2. *Under the blanket assumptions [Assumptions 6.1](#), [6.3](#) and [D.1](#) and with the additional [Assumption 6.5](#), for any threshold $\delta \in (0, 1)$, define*

$$\underline{\lambda}(\rho) = \begin{cases} \frac{\lambda_0^*}{32\rho} & \text{if } \rho \leq \rho_{\text{thres}} = \min \left(\frac{\lambda_0^*}{32(\lambda_1 + L_2)}, \frac{\mu^*(\lambda_0^*)^2}{4096L_2} \right) \\ \min \left(\lambda_2, \frac{\rho_c^2(0,0) - \rho^2}{2L} \right) & \text{otherwise} \end{cases} \quad (\text{D.5.8})$$

$$\bar{\lambda}(\rho) = \frac{\lambda_0^*}{2\rho}. \quad (\text{D.5.9})$$

Assume that

$$\rho \geq \frac{8192}{\sqrt{n}\mu^*(\lambda_0^*)^2} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \left(\tilde{F}(0) + \bar{M}(\rho) \right) \sqrt{1 + \log \frac{4}{\delta}} \right), \quad (\text{D.5.10})$$

and that

$$\rho^2 \leq \rho_c^2(0, 0) - \frac{2B(\delta)}{\sqrt{n}}. \quad (\text{D.5.11})$$

Then, with probability at least $1 - \frac{\delta}{2}$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) = \inf_{\underline{\lambda}(\rho) \leq \lambda \leq \bar{\lambda}(\rho)} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, 0)] \quad (\text{D.5.12})$$

and when $\rho^2 \geq \bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0)$, with probability $1 - \delta$, it holds,

$$\widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathcal{R}_{\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0)}(f). \quad (\text{D.5.13})$$

Furthermore, with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \geq \sup \left\{ \mathbb{E}_Q[f] : Q \in \mathcal{P}(\Xi), W_2^2(P, Q) \leq \rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0) \right\}. \quad (\text{D.5.14})$$

Proof. This result is a consequence of [Propositions D.3.1](#) and [D.4.1](#) both applied with $\delta \leftarrow \delta/4$ and of [Proposition D.2.1](#). Note that the upper-bound on the dual variable given by [Proposition D.3.1](#) holds for any ρ since the optimal dual variable is non-increasing as a function of ρ . ■

Proof of Theorem D.1. The proof consists in simplifying both the assumptions and the result of Lemma D.5.2.

We begin by showing that $\underline{\lambda}(\rho)$ can always be lower-bounded by a quantity proportional to $1/\rho$. Indeed, by definition of $\underline{\lambda}(\rho)$, (D.5.8) in Lemma D.5.2, and using that ρ is in particular less than $\frac{\rho_c}{2}$, it holds that,

$$\underline{\lambda}(\rho) \geq \frac{1}{\rho} \min\left(\frac{\lambda_0^*}{32}, \rho_{\text{thres}} \lambda_2, \frac{3\rho_c^2 \rho_{\text{thres}}}{8L}\right) \quad (\text{D.5.15})$$

Let us now turn our attention to the condition $\rho^2 \geq \bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0, 0)$, whose RHS was defined by (D.2.1) in Appendix D.2. We have that, by definition (Assumption D.1), $\sup_{0 < \rho \leq \rho_c} \tilde{F}(\lambda_0^*/(32\rho)) = \tilde{F}(\lambda_0^*/(32\rho_c)) < +\infty$ and,

$$\sup_{0 < \rho \leq \rho_c} a\left(\underline{\lambda}_n(\rho), \bar{\lambda}(\rho), 0, 0\right) \leq \sup_{0 < \rho' \leq \rho_{\text{thres}}} a\left(\frac{1}{\rho'} \min\left(\frac{\lambda_0^*}{32}, \rho_{\text{thres}} \lambda_2, \frac{3\rho_c^2 \rho_{\text{thres}}}{8L}\right), \frac{\lambda_0^*}{2\rho'}, 0, 0\right) \quad (\text{D.5.16})$$

$$= \bar{a} < +\infty, \quad (\text{D.5.17})$$

by definition and non-decreasingness of a in its first argument (see Corollary D.1.5) and (D.5.15). Hence, the following bound holds

$$\bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), \varepsilon, \sigma) \quad (\text{D.5.18})$$

$$\leq \frac{117}{\sqrt{n} \underline{\lambda}(\rho)} \left(\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \max\left(\tilde{F}\left(\frac{\lambda_0^*}{32\rho_c}\right), \bar{a}\right) \left(1 + \sqrt{\log \frac{1}{\delta}}\right) \right) \quad (\text{D.5.19})$$

$$\leq \rho_n \rho, \quad (\text{D.5.20})$$

where we plugged (D.5.15).

Finally, since ρ is in particular bounded by ρ_c , the condition (D.5.10) is implied by

$$\rho \geq \frac{8192}{\sqrt{n} \mu^*(\lambda_0^*)^2} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \left(\tilde{F}(0) + \bar{M}(\rho_c)\right) \sqrt{1 + \log \frac{4}{\delta}} \right), \quad (\text{D.5.21})$$

with $\bar{M}(\rho_c) < +\infty$ by definition (Proposition D.3.1). $\blacksquare$

Theorem D.3 (Extended version of Theorem 6.2). *Under Assumptions 6.1, 6.3 and D.1, for any $\delta \in (0, 1)$ and $n \geq 1$, if*

$$\max\left(\rho_n, \frac{8192}{\sqrt{n} \mu^*(\lambda_0^*)^2} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \left(\tilde{F}(0) + \bar{M}(\rho_c)\right) \sqrt{1 + \log \frac{2}{\delta}} \right)\right) \leq \rho \quad (\text{D.5.22})$$

$$\text{and} \quad \rho \leq \min\left(\rho_{\text{thres}}, \frac{\rho_c}{2} - \frac{96\mathcal{I}(\mathcal{F}, D) + 4\sqrt{C^* \log 1/\delta}}{\sqrt{n}}\right). \quad (\text{D.5.23})$$

where

$$\rho_{\text{thres}} := \min\left(\frac{\lambda_0^*}{8(\lambda_1 + L_2)}, \frac{\mu^*(\lambda_0^*)^2}{4096L_2}\right) \quad (\text{D.5.24})$$

$$\bar{a} := \sup_{0 < \rho' \leq \rho_{\text{thres}}} a\left(\frac{\lambda_0^*}{32\rho'}, \frac{\lambda_0^*}{2\rho'}, 0, 0\right) \quad (\text{D.5.25})$$

$$\rho_n := \frac{3744 \left(\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \max\left(\tilde{F}\left(\frac{\lambda_0^*}{32\rho_{\text{thres}}}\right), \bar{a}\right) \left(1 + \sqrt{\log \frac{1}{\delta}}\right) \right)}{\sqrt{n}\lambda_0^*}, \quad (\text{D.5.26})$$

then, with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathbb{E}_{\xi \sim Q} [f(\xi)] \quad \text{for all } Q \text{ such that } W_2^2(P, Q) \leq \rho(\rho - \rho_n). \quad (\text{D.5.27})$$

In particular, with probability $1 - \delta$, we have

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathbb{E}_{\xi \sim P} [f(\xi)]. \quad (\text{D.5.28})$$

The proof of [Theorem D.3](#) leverages results from the previous sections, combined in [Lemma D.5.4](#).

Lemma D.5.4. *Under the blanket assumptions [Assumptions 6.1](#), [6.3](#) and [D.1](#), for any threshold $\delta \in (0, 1)$, define*

$$\underline{\lambda}(\rho) = \frac{\lambda_0^*}{32\rho}, \quad \bar{\lambda}(\rho) = \frac{\lambda_0^*}{2\rho}. \quad (\text{D.5.29})$$

Assume that

$$\rho \geq \frac{8192}{\sqrt{n}\mu^*(\lambda_0^*)^2} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \left(\tilde{F}(0) + \overline{M}(\rho) \right) \sqrt{1 + \log \frac{2}{\delta}} \right), \quad (\text{D.5.30})$$

and that

$$\rho^2 \leq \min \left(\rho_c^2(0, 0) - \frac{2B(\delta)}{\sqrt{n}}, \min \left(\frac{\lambda_0^*}{32(\lambda_1 + L_2)}, \frac{\mu^*(\lambda_0^*)^2}{4096L_2} \right)^2 \right). \quad (\text{D.5.31})$$

Then, with probability at least $1 - \frac{\delta}{2}$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) = \inf_{\underline{\lambda}(\rho) \leq \lambda \leq \bar{\lambda}(\rho)} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, 0)] \quad (\text{D.5.32})$$

and when $\rho^2 \geq \bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0)$, with probability $1 - \delta$, it holds,

$$\widehat{\mathcal{R}}_{\rho^2}(f) \geq \mathcal{R}_{\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0)}(f). \quad (\text{D.5.33})$$

Furthermore, with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \geq \sup \left\{ \mathbb{E}_Q[f] : Q \in \mathcal{P}(\Xi), W_2^2(P, Q) \leq \rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0) \right\}. \quad (\text{D.5.34})$$

Proof. This result follows directly from [Proposition D.3.1](#) that we invoke with $\delta \leftarrow \delta/2$ and of [Proposition D.2.1](#). $\blacksquare$

Proof of [Theorem D.3](#). The proof consists in simplifying both the assumptions and the result of [Lemma D.5.4](#) and follows the same structure as the proof of [Theorem D.1](#).

We begin by examining the condition $\rho^2 \geq \bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0, 0)$, whose RHS was defined by (D.2.1) in [Appendix D.2](#). We have that, by definition ([Assumption D.1](#)), $\sup_{0 < \rho \leq \rho_c} \tilde{F}(\lambda_0^*/(32\rho)) = \tilde{F}(\lambda_0^*/(32\rho_c)) < +\infty$ and,

$$\sup_{0 < \rho \leq \rho_{\text{thres}}} a\left(\underline{\lambda}_n(\rho), \bar{\lambda}(\rho), 0, 0\right) = \sup_{0 < \rho' \leq \rho_{\text{thres}}} a\left(\frac{\lambda_0^*}{32\rho'}, \frac{\lambda_0^*}{2\rho'}, 0, 0\right) = \bar{a} < +\infty \quad (\text{D.5.35})$$

by definition (see [Corollary D.1.5](#)). Hence, we have that

$$\bar{\rho}_n^2(\delta, \underline{\lambda}(\rho), \bar{\lambda}(\rho), 0, 0) \quad (\text{D.5.36})$$

$$\leq \frac{117}{\sqrt{n}\underline{\lambda}(\rho)} \left(\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \max\left(\tilde{F}\left(\frac{\lambda_0^*}{32\rho_c}\right), \bar{a}\right) \left(1 + \sqrt{\log \frac{1}{\delta}}\right) \right) \quad (\text{D.5.37})$$

$$\leq \rho_n \rho, \quad (\text{D.5.38})$$

by definition of ρ_n and with $117 \times 32 = 3744$.

Finally, [m \(D.5.30\)](#) is implied by

$$\rho \geq \frac{8192}{\sqrt{n}\mu^*(\lambda_0^*)^2} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \left(\tilde{F}(0) + \bar{M}(\rho_{\text{thres}}) \right) \sqrt{1 + \log \frac{2}{\delta}} \right), \quad (\text{D.5.39})$$

since $\rho \leq \rho_{\text{thres}}$ and $\bar{M}(\rho_{\text{thres}}) < +\infty$ by definition ([Proposition D.3.1](#)). $\blacksquare$

D.5.2 Regularized WDRO case

Theorem D.5 (Extended version of [Theorem 6.6](#)). *For $\sigma = \sigma_0\rho$ with $\sigma_0 > 0$, $\varepsilon = \varepsilon_0\rho$ with $\varepsilon_0 > 0$ such that $\varepsilon_0/\sigma_0^2 \leq \lambda_0^*/8$, and for any $\delta \in (0, 1)$ and $n \geq 1$, define,*

$$\rho_c := \inf\{\rho_c(\varepsilon_0\rho', \sigma_0\rho') : \rho_{\text{thres}} \leq \rho' \leq \rho_c(\varepsilon_0\rho_{\text{thres}}, \sigma_0\rho_{\text{thres}})\} \quad (\text{D.5.40})$$

and

$$\rho_{\text{thres}} := \min\left(\frac{\varepsilon_1}{\varepsilon_0}, \frac{\lambda_0^*}{32(\lambda_1 + L_2)}, \frac{\mu^*(\lambda_0^*)^2}{4096L_2}, \sqrt{\frac{c_2^3\lambda_0^*}{8\varepsilon_0}} \left(\log\left(\frac{4096\varepsilon_0c_1}{\mu^*(\lambda_0^*)^2}\right) \right)_+^{-\frac{3}{2}}\right) \quad (\text{D.5.41})$$

$$\bar{\text{Var}} := \sup_{\rho_{\text{thres}} \leq \rho' \leq \rho_c} \text{Var}(\varepsilon_0\rho', \sigma_0\rho') \quad (\text{D.5.42})$$

$$\bar{C}_{\mathcal{F}} := \sup_{\rho_{\text{thres}} \leq \rho' \leq \rho_c} C_{\mathcal{F}}(\varepsilon_0\rho', \sigma_0\rho') \quad (\text{D.5.43})$$

$$\bar{a} := \sup_{0 < \rho' \leq \rho_c} a\left(\frac{1}{\rho'}, \min\left(\frac{\lambda_0^*}{32}, \frac{\varepsilon_0\rho_{\text{thres}}^2\rho_c^2}{4\bar{\text{Var}}}\right), \max\left(\frac{\lambda_0^*}{2\rho'}, \frac{12\varepsilon_0\rho_c \log(2 \times 6^{d/2})}{R^2}, \frac{e^{\frac{\sup_{f \in \mathcal{F}} \|f\|_\infty}{\varepsilon_0\rho_{\text{thres}}}}}{\rho_{\text{thres}}}\right), \varepsilon_0\rho', \sigma_0\rho'\right) \quad (\text{D.5.44})$$

$$\rho_n := \frac{117 \left(\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \max\left(\tilde{F}\left(\frac{\lambda_0^*}{32\rho_c}\right), \bar{a}\right) \left(1 + \sqrt{\log \frac{1}{\delta}}\right) \right)}{\sqrt{n} \min\left(\frac{\lambda_0^*}{32}, \frac{\varepsilon_0\rho_{\text{thres}}^2\rho_c^2}{4\bar{\text{Var}}}\right)}; \quad (\text{D.5.45})$$

when

$$\max\left(\rho_n, \frac{8192}{\mu^*(\lambda_0^*)^2 \sqrt{n}} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + (\tilde{F}(0) + \overline{M}(\rho_c)) \sqrt{\log \frac{4}{\delta}}\right), \frac{384\sqrt{\text{Var}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}}{\varepsilon_0 \rho_c^2 \sqrt{n}}\right) \leq \rho \quad (\text{D.5.46})$$

$$\rho \leq \frac{\rho_c}{2} - \frac{384\sqrt{\text{Var}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}}{\varepsilon_0 \rho_c^2 \sqrt{n}} \quad (\text{D.5.47})$$

$$\text{and } \rho_c \geq \max\left(\left(\frac{192\sqrt{\text{Var}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}}{\varepsilon_0 \sqrt{n}}\right)^{1/3}, 2\sqrt{C_{\mathcal{F}}}\left(\frac{\log \frac{4}{\delta}}{2n}\right)^{1/4}\right), \quad (\text{D.5.48})$$

then, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \mathbb{E}_{\xi \sim Q} [f(\xi)] \quad \text{for all } Q \text{ such that } W_{2, \tau(\rho)}^2(P, Q) \leq \rho(\rho - \rho_n), \quad (\text{D.5.49})$$

where $\tau(\rho) \leq \frac{\varepsilon \rho}{\min\left(\frac{\lambda_0^*}{32}, \frac{\varepsilon_0 \rho_{\text{thres}}^2 \rho_c^2}{4\text{Var}}\right)}$. Furthermore, when $\sigma_0 \leq 1$ and $\sigma \leq \sigma_1$ (defined in [Proposition D.1.2](#)), with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \mathbb{E}_{\xi \sim P} \mathbb{E}_{\zeta \sim \pi_{\sigma}(\cdot|\xi)} [f(\zeta)]. \quad (\text{D.5.50})$$

The proof of [Theorem D.5](#) relies on [Lemma D.5.6](#) that makes the regularized Wasserstein distance appear. It also uses [Lemma D.5.7](#), to guarantee that a smoothed version of the true distribution is inside the right neighborhood.

Lemma D.5.6. Fix a confidence threshold $\delta \in (0, 1)$, take $\varepsilon = \varepsilon_0 \rho$, $\sigma = \sigma_0 \rho$ with ε_0 and σ_0 positive constants satisfying $\varepsilon_0 / \sigma_0^2 \leq \lambda_0^* / 8$ and, define $\underline{\lambda}_n(\rho)$ and $\overline{\lambda}(\rho)$ as functions of ρ by

• If

$$\rho \leq \min\left(\frac{\varepsilon_1}{\varepsilon_0}, \frac{\lambda_0^*}{32(\lambda_1 + L_2)}, \frac{\mu^*(\lambda_0^*)^2}{4096L_2}, \sqrt{\frac{c_2^3 \lambda_0^*}{8\varepsilon_0}} \left(\log\left(\frac{4096\varepsilon_0 c_1}{\mu^*(\lambda_0^*)^2}\right)\right)_+^{-\frac{3}{2}}\right), \quad (\text{D.5.51})$$

$$\text{then } \underline{\lambda}_n(\rho) = \frac{\lambda_0^*}{32\rho} \text{ and } \overline{\lambda}(\rho) = \frac{\lambda_0^*}{2\rho},$$

• Otherwise,

$$\underline{\lambda}_n(\rho) = \frac{\varepsilon_0 \rho}{\text{Var}(\varepsilon_0 \rho, \sigma_0 \rho)} \left(\rho_c(\varepsilon_0 \rho, \sigma_0 \rho)^2 - \rho^2 - \left(\frac{48\sqrt{\text{Var}(\varepsilon_0 \rho, \sigma_0 \rho)} \mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\varepsilon_0 \rho \sqrt{n}} + 2C_{\mathcal{F}}(\varepsilon_0 \rho, \sigma_0 \rho) \sqrt{\frac{\log}{2n}} \right) \right) \quad (\text{D.5.52})$$

$$\overline{\lambda}(\rho) = \max\left(\frac{12\varepsilon_0 \rho}{R^2} \log(2 \times 6^{d/2}), e^{\frac{\sup_{f \in \mathcal{F}} \|f\|_\infty}{\varepsilon_0 \rho}} \frac{\varepsilon_0}{\rho}\right). \quad (\text{D.5.53})$$

Assume that

$$\rho \geq \frac{8192}{\sqrt{n} \mu^*(\lambda_0^*)^2} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + (\tilde{F}(0) + \overline{M}(\rho)) \sqrt{1 + \log \frac{4}{\delta}}\right), \quad (\text{D.5.54})$$

Then, with probability at least $1 - \frac{\delta}{2}$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda}_n(\rho) \leq \lambda \leq \bar{\lambda}(\rho)} \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.5.55})$$

and when $\rho^2 \geq \bar{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \bar{\lambda}(\rho), \varepsilon, \sigma)$, with probability $1 - \delta$, it holds,

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \mathcal{R}_{\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \bar{\lambda}(\rho), \varepsilon, \sigma)}^\varepsilon(f). \quad (\text{D.5.56})$$

Furthermore, with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \sup \left\{ \mathbb{E}_Q[f] : Q \in \mathcal{P}(\Xi), W_{2, \tau(\rho)}^2(P, Q) \leq \rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \bar{\lambda}(\rho), \varepsilon, \sigma) \right\}, \quad (\text{D.5.57})$$

with $\tau(\rho) := \frac{\varepsilon_0 \rho}{\underline{\lambda}_n(\rho)}$.

Proof. The first part of this result is a consequence of the combination of [Propositions D.3.1](#) and [D.4.3](#), both applied with $\delta \leftarrow \delta/4$, and of [Proposition D.2.1](#). For the second part, note that [Remark 12](#) implies that the above argument actually gives the slightly stronger result: with probability $1 - \delta$, for any $f \in \mathcal{F}$,

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \inf_{\underline{\lambda}_n(\rho) \leq \lambda \leq \bar{\lambda}(\rho)} \lambda(\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \bar{\lambda}(\rho), \varepsilon, \sigma)) + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.5.58})$$

Next, take $Q \in \mathcal{P}(\Xi)$ such that $W_{2, \tau(\rho)}^2(P, Q) \leq \rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \varepsilon, \sigma)$. With a similar argument as in the proof of [Proposition D.2.1](#), we get that

$$\widehat{\mathcal{R}}_{\rho^2}^\varepsilon(f) \geq \inf_{\underline{\lambda}_n(\rho) \leq \lambda \leq \bar{\lambda}(\rho)} \lambda(\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \bar{\lambda}(\rho), \varepsilon, \sigma)) + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.5.59})$$

$$= \mathbb{E}_Q[f] + \inf_{\underline{\lambda}_n(\rho) \leq \lambda \leq \bar{\lambda}(\rho)} \lambda(\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \bar{\lambda}(\rho), \varepsilon, \sigma)) - \{\mathbb{E}_Q[f] - \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]\} \quad (\text{D.5.60})$$

$$= \mathbb{E}_Q[f] + \inf_{\underline{\lambda}_n(\rho) \leq \lambda \leq \bar{\lambda}(\rho)} \lambda(\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \bar{\lambda}(\rho), \varepsilon, \sigma)) - \sup_{f' \in \mathcal{F}} \{\mathbb{E}_Q[f'] - \mathbb{E}_{\xi \sim P} [\phi(f', \xi, \lambda, \varepsilon, \sigma)]\}. \quad (\text{D.5.61})$$

We now proceed to show, and this will conclude the proof, that

$$\sup_{f \in \mathcal{F}} \{\mathbb{E}_Q[f] - \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]\} \leq \lambda W_{2, \tau(\rho)}^2(P, Q), \quad (\text{D.5.62})$$

for $\lambda \geq \underline{\lambda}_n(\rho)$.

Indeed,

$$\sup_{f \in \mathcal{F}} \{\mathbb{E}_Q[f] - \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]\} \leq \sup_{f \in \mathcal{C}(\Xi)} \{\mathbb{E}_Q[f] - \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]\} \quad (\text{D.5.63})$$

$$= \sup_{f \in \mathcal{C}(\Xi)} \left\{ \mathbb{E}_Q[f] - \mathbb{E}_{\xi \sim P} \left[\log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot | \xi)} \left[e^{\frac{f(\zeta) - \lambda \|\xi - \zeta\|^2 / 2}{\varepsilon}} \right] \right) \right] \right\} \quad (\text{D.5.64})$$

$$= \lambda \sup_{f \in \mathcal{C}(\Xi)} \left\{ \mathbb{E}_Q[f] - \mathbb{E}_{\xi \sim P} \left[\log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[e^{\frac{f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2}{\varepsilon/\lambda}} \right] \right) \right] \right\}. \quad (\text{D.5.65})$$

where we performed the change of variable $f \leftarrow f/\lambda$. We now show the following equality that will allow us to rewrite the RHS of (D.5.65).

$$-\mathbb{E}_{\xi \sim P} \left[\log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[e^{\frac{f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2}{\varepsilon/\lambda}} \right] \right) \right] = \sup_{g \in \mathcal{C}(\Xi)} \mathbb{E}_P[g] - \frac{\varepsilon}{\lambda} \left(\mathbb{E}_{(\xi, \zeta) \sim \pi_\sigma} \left[e^{\frac{g(\xi) + f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2}{\varepsilon/\lambda}} \right] - 1 \right). \quad (\text{D.5.66})$$

Solving the optimality condition of the concave problem of the RHS of (D.5.66) gives that its maximum is reached for

$$g(\xi) = -\log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[e^{\frac{f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2}{\varepsilon/\lambda}} \right] \right) \quad (\text{D.5.67})$$

so that (D.5.66) holds. Hence, we get that

$$\sup_{f \in \mathcal{C}(\Xi)} \left\{ \mathbb{E}_Q[f] - \mathbb{E}_{\xi \sim P} \left[\log \left(\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[e^{\frac{f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2}{\varepsilon/\lambda}} \right] \right) \right] \right\} \quad (\text{D.5.68})$$

$$= \sup_{f, g \in \mathcal{C}(\Xi)} \left\{ \mathbb{E}_Q[f] + \mathbb{E}_P[g] - \frac{\varepsilon}{\lambda} \left(\mathbb{E}_{(\xi, \zeta) \sim \pi_\sigma} \left[e^{\frac{g(\xi) + f(\zeta) - \frac{1}{2} \|\xi - \zeta\|^2}{\varepsilon/\lambda}} \right] - 1 \right) \right\} \quad (\text{D.5.69})$$

$$= W_{2, \varepsilon/\lambda}^2(P, Q), \quad (\text{D.5.70})$$

by the duality formula for regularized OT (Peyré et al., 2019)¹.

Combining this equality with the bound of (D.5.65) gives

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_Q[f] - \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \right\} \leq \lambda W_{2, \varepsilon/\lambda}^2(P, Q), \quad (\text{D.5.71})$$

which yields the result since $W_{2, \tau}^2(P, Q)$ is non-decreasing in τ . $\blacksquare$

Lemma D.5.7. In the setting of Theorem D.5, when Q_σ denotes the second marginal of

$$P(d\xi)\pi_\sigma(d\zeta|\xi), \quad (\text{D.5.72})$$

and when $\sigma \leq \sigma_1$, it holds

$$W_{2, \tau(\rho)}^2(P, Q_\sigma) \leq \sigma^2. \quad (\text{D.5.73})$$

Proof. Consider the transport plan $\pi = P(d\xi)\pi_\sigma(d\zeta|\xi)$. To show this lemma, it suffices to prove that

$$\mathbb{E}_\pi[d^2] + \tau(\rho) \text{KL}(\pi|\pi) = \mathcal{O}(\sigma^2), \quad (\text{D.5.74})$$

¹ To get this exact result for a regularization w.r.t. an arbitrary measure, one can readily combine Paty and Cuturi (2020a, Cor. 1) and Feydy et al. (2019, Prop. 7). Also, note that we essentially reproved the semi-duality formula of Genevay et al. (2016, Prop. 2.1) except that the regularization is taken w.r.t. a general measure.

i.e., that $\mathbb{E}_\pi[d^2] = \mathcal{O}(\sigma^2)$. Let us first fix $\xi \in \text{supp } P$ and consider $\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right]$, which is equal to

$$\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] = \frac{\int_{\Xi} \frac{1}{2} \|\xi - \zeta\|^2 e^{-\frac{\|\xi - \zeta\|^2}{2\sigma^2}} d\zeta}{\int_{\Xi} e^{-\frac{\|\xi - \zeta\|^2}{2\sigma^2}} d\zeta}. \quad (\text{D.5.75})$$

The numerator can be upper-bounded as follows:

$$\int_{\Xi} \frac{1}{2} \|\xi - \zeta\|^2 e^{-\frac{\|\xi - \zeta\|^2}{2\sigma^2}} d\zeta \leq \int_{\mathbb{R}^d} \frac{1}{2} \|\xi - \zeta\|^2 e^{-\frac{\|\xi - \zeta\|^2}{2\sigma^2}} d\zeta = (2\pi\sigma^2)^{d/2} \frac{\sigma^2}{2}. \quad (\text{D.5.76})$$

For the denominator, we have seen in the proof of [Lemma D.1.4](#), and more precisely [\(D.1.24\)](#), that

$$\left(\int_{\Xi} e^{-\frac{\|\xi - \zeta\|^2}{2\sigma^2}} d\zeta \right)^{-1} \leq \frac{2}{(2\pi\sigma^2)^{d/2}}, \quad (\text{D.5.77})$$

when $\sigma \leq \sigma_1$. Hence, we have the bound

$$\mathbb{E}_{\zeta \sim \pi_\sigma(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \leq \sigma^2, \quad (\text{D.5.78})$$

and integrating w.r.t. $\xi \sim P$ yields the result. $\blacksquare$

Proof of Theorem D.5. Since we will only consider radii in particular bounded by ρ_c , the condition [\(D.5.54\)](#) is implied by

$$\rho \geq \frac{8192}{\mu^*(\lambda_0^*)^2 \sqrt{n}} \left(12\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \left(\widetilde{F}(0) + \overline{M}(\rho_c) \right) \sqrt{\log \frac{4}{\delta}} \right), \quad (\text{D.5.79})$$

with $\overline{M}(\rho_c) < +\infty$.

We now show that $\underline{\lambda}_n(\rho)$ can always be lower-bounded by a quantity proportional to $1/\rho$, i.e., that

$$\underline{\lambda}_n(\rho) \geq \frac{1}{\rho} \min \left(\frac{\lambda_0^*}{32}, \frac{\varepsilon_0 \rho_{\text{thres}}^2 \rho_c^2}{4\text{Var}} \right). \quad (\text{D.5.80})$$

Let us discuss separately the cases where $\rho \leq \rho_{\text{thres}}$ holds or not.

- When $\rho \leq \rho_{\text{thres}}$, [\(D.5.80\)](#) holds by definition of $\underline{\lambda}_n(\rho)$.
- When $\rho > \rho_{\text{thres}}$, by definition, $\underline{\lambda}_n(\rho)$ is lower bounded as

$$\underline{\lambda}_n(\rho) \geq \frac{\varepsilon_0 \rho}{\text{Var}} \left(\rho_c^2 - \rho^2 - \left(\frac{48\sqrt{\text{Var}}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\varepsilon_0 \rho \sqrt{n}} + 2\overline{C}_{\mathcal{F}} \sqrt{\frac{\log \frac{4}{\delta}}{2n}} \right) \right). \quad (\text{D.5.81})$$

Applying [Lemma D.7.9](#) with $\bar{\rho} \leftarrow \frac{\rho_c}{2}$ and $c \leftarrow \frac{48\sqrt{\text{Var}}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\varepsilon_0 \sqrt{n}}$, we obtain that, when

$$\rho_c \geq \left(\frac{192\sqrt{\text{Var}}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}{\varepsilon_0 \sqrt{n}} \right)^{1/3} \quad (\text{D.5.82})$$

and

$$\frac{384\sqrt{\text{Var}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}}{\varepsilon_0\rho_c^2\sqrt{n}} \leq \rho \leq \frac{\rho_c}{2} - \frac{384\sqrt{\text{Var}\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty)}}{\varepsilon_0\rho_c^2\sqrt{n}}, \quad (\text{D.5.83})$$

the following lower-bound holds,

$$\underline{\lambda}_n(\rho) \geq \frac{\varepsilon_0\rho}{\text{Var}} \left(\frac{3\rho_c^2}{4} - 2\overline{C}_{\mathcal{F}} \sqrt{\frac{\log \frac{4}{\delta}}{2n}} \right) \geq \frac{\varepsilon_0\rho\rho_c^2}{4\text{Var}} \geq \frac{\varepsilon_0\rho_{\text{thres}}^2\rho_c^2}{4\text{Var}\rho}, \quad (\text{D.5.84})$$

where we used successively that $\frac{\rho_c^2}{2} \geq 2\overline{C}_{\mathcal{F}} \sqrt{\frac{\log \frac{4}{\delta}}{2n}}$ and $\rho \geq \rho_{\text{thres}}$. This concludes the proof of (D.5.80). Note that it implies the bound on $\tau(\rho)$ in the statement.

Let us finally turn our attention to the condition $\rho^2 \geq \overline{\rho}_n^2(\delta, \underline{\lambda}_n(\rho), \overline{\lambda}(\rho), \varepsilon, \sigma)$. Since $\sup_{0 < \rho \leq \rho_c} \tilde{F}(\lambda_0^*/(32\rho)) = \tilde{F}(\lambda_0^*/(32\rho_c)) < +\infty$ by definition (Assumption D.1) and

$$\sup_{0 < \rho' \leq \rho_c} a(\underline{\lambda}_n(\rho'), \overline{\lambda}(\rho'), \varepsilon_0\rho', \varepsilon_0\sigma) \quad (\text{D.5.85})$$

$$\leq \sup_{0 < \rho' \leq \rho_c} a\left(\frac{1}{\rho'} \min\left(\frac{\lambda_0^*}{32}, \frac{\varepsilon_0\rho_{\text{thres}}^2\rho_c^2}{4\text{Var}}\right), \max\left(\frac{\lambda_0^*}{2\rho'}, \frac{12\varepsilon_0\rho_c \log(2 \times 6^{d/2})}{R^2}, \frac{e^{\frac{\sup_{f \in \mathcal{F}} \|f\|_\infty}{\varepsilon_0\rho_{\text{thres}}}} \varepsilon_0}{\rho_{\text{thres}}}\right), \varepsilon_0\rho', \sigma_0\rho'\right) \quad (\text{D.5.86})$$

$$= \overline{a} < +\infty \quad (\text{D.5.87})$$

$$(\text{D.5.88})$$

where we used the monotonicity properties of a (Corollary D.1.5) and (D.5.80).

In conclusion, along with (D.5.80), we obtain that

$$\overline{\rho}_n^2(\delta, \underline{\lambda}(\rho), \overline{\lambda}(\rho), \varepsilon, \sigma) \quad (\text{D.5.89})$$

$$\leq \frac{117}{\sqrt{n}\underline{\lambda}(\rho)} \left(\mathcal{I}(\mathcal{F}, \|\cdot\|_\infty) + \max\left(\tilde{F}\left(\frac{\lambda_0^*}{32\rho_c}\right), \overline{a}\right) \left(1 + \sqrt{\log \frac{1}{\delta}}\right) \right) \quad (\text{D.5.90})$$

$$\leq \rho_n \rho, \quad (\text{D.5.91})$$

by definition of ρ_n . The last part of the statement then follows from Lemma D.5.7. $\blacksquare$

D.6 UPPER-BOUND ON THE EMPIRICAL ROBUST RISK

In this section we prove Theorem 6.3 that complements the main results by providing both a lower and an upper bound on the empirical robust risk. In view of the previous section, the missing part is the upper-bound, that we establish in this section.

The proof of the upper-bound is similar to the proof of our main results, yet simpler. Indeed, the bounds on the dual variable are required for the true distribution P , which is fixed, instead of the empirical distribution $\widehat{P}_n$. We slightly modify our main concentration result (Proposition D.2.1) in Proposition D.6.1. We simplify our bounds on the dual multiplier when the radius is close to the critical radius (Propositions D.4.1 and D.4.3) in Propositions D.6.2 and D.6.5.

D.6.1 From empirical to true risk

Proposition D.6.1. For $\rho > 0$, $\varepsilon \geq 0$, $\sigma > 0$ and $\delta \in (0, 1)$, assume that there is some $0 < \underline{\lambda} \leq \bar{\lambda} < +\infty$ such that,

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \lambda \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] . \quad (\text{D.6.1})$$

Then, when $\rho^2 \geq \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)$, with probability $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)}^\varepsilon(f) \leq \mathcal{R}_{\rho^2}^\varepsilon(f) . \quad (\text{D.6.2})$$

Proof. This proof closely mimics the one of [Proposition D.2.1](#) but switches the roles of P and $\widehat{P}_n$. First, note that by following the proof of [Lemma D.2.3](#) with and replacing P by $\widehat{P}_n$ and *vice versa* (and using statement (a) of [Lemma D.7.2](#) instead of (b)) yields the following

$$\sup_{(f, \lambda) \in \mathcal{F} \times [\underline{\lambda}, \bar{\lambda}]} \left\{ \frac{\mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]}{\lambda} \right\} \leq \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) . \quad (\text{D.6.3})$$

We can now follow the last part of the proof of [Proposition D.2.1](#). On the event above, for any $f \in \mathcal{F}$,

$$\mathcal{R}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \right\} \quad (\text{D.6.4})$$

$$= \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \lambda \frac{\mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)]}{\lambda} \right\} \quad (\text{D.6.5})$$

$$\geq \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \lambda \sup_{\underline{\lambda} \leq \lambda' \leq \bar{\lambda}} \frac{\mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda', \varepsilon, \sigma)] - \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda', \varepsilon, \sigma)]}{\lambda'} \right\} \quad (\text{D.6.6})$$

$$\geq \inf_{\underline{\lambda} \leq \lambda \leq \bar{\lambda}} \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \widehat{P}_n} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] - \lambda \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma) \right\} \quad (\text{D.6.7})$$

$$\geq \widehat{\mathcal{R}}_{\rho^2 - \bar{\rho}_n^2(\delta, \underline{\lambda}, \bar{\lambda}, \varepsilon, \sigma)}^\varepsilon(f) . \quad (\text{D.6.8})$$

■

D.6.2 Standard WDRO case

Proposition D.6.2. Let [Assumption 6.5](#) hold and fix a threshold $\delta \in (0, 1)$. Assume that $\rho^2 \leq \rho_c^2(0, 0)$. Then, we have,

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_{\rho^2}(f) = \inf_{\underline{\lambda} \leq \lambda} \lambda \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, 0)] \quad (\text{D.6.9})$$

where the dual bound $\underline{\lambda}$ is defined as

$$\underline{\lambda} := \min \left(\lambda_2, \frac{\rho_c^2(0, 0) - \rho^2}{L} \right), \quad (\text{D.6.10})$$

and L is defined in [Proposition D.4.1](#).

Proof. Let $0 \leq \lambda \leq \underline{\lambda}$. By [Lemma D.4.2](#) and the dominated convergence theorem, one has that,

$$\partial_\lambda \psi_\rho(\lambda) = \rho^2 - \mathbb{E}_{\xi \sim P} \left[\min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max_{\Xi} f - \frac{\lambda}{2} \|\xi - \cdot\|^2 \right\} \right] \quad (\text{D.6.11})$$

$$\leq \rho^2 - \left(1 - \frac{16\lambda}{\mu} \right) \mathbb{E}_{\xi \sim P} \left[\min \left\{ \frac{1}{2} \|\xi - \zeta\|^2 : \zeta \in \arg \max_{\Xi} f \right\} \right] \quad (\text{D.6.12})$$

$$\leq \rho^2 - \rho_c^2(0, 0) + L\lambda, \quad (\text{D.6.13})$$

which is non-negative by definition of $\underline{\lambda}$ and thus concludes the proof. $\blacksquare$

We can now state analogues of [Theorems D.1](#) and [D.3](#). Note that the bounds $\underline{\lambda}(\rho)$ that we obtained in this section are better than the ones we got in the main proof. For the sake of simplicity, we give up this additional precision and use the same bounds as in [Theorems D.1](#) and [D.3](#).

Corollary D.6.3. *In the same setting as [Theorem D.1](#), with probability $1 - \delta$, it holds,*

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_{\rho^2}(f) \geq \widehat{\mathcal{R}}_{\rho(\rho-\rho_n)}(f). \quad (\text{D.6.14})$$

Proof. This result is obtained as a combination of [Propositions D.3.1](#), [D.6.1](#) and [D.6.2](#), which gives the desired result with probability at least $1 - \frac{\delta}{2}$ and *a fortiori* $1 - \delta$. $\blacksquare$

Corollary D.6.4. *In the same setting as [Theorem D.3](#), with probability $1 - \delta$, it holds,*

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_{\rho^2}(f) \geq \widehat{\mathcal{R}}_{\rho(\rho-\rho_n)}(f). \quad (\text{D.6.15})$$

Proof. This result follows by combining [Propositions D.3.1](#) and [D.6.1](#), which gives the desired result with probability at least $1 - \frac{\delta}{2}$ and *a fortiori* $1 - \delta$. $\blacksquare$

To conclude, in the context of [Theorem D.1](#) (resp. [Theorem D.3](#)), [Corollary D.6.3](#) (resp. [Corollary D.6.4](#)) with $\rho \leftarrow \rho + \rho_n$ yields, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho(\rho+\rho_n)}(f) \leq \mathcal{R}_{(\rho+\rho_n)^2}(f), \quad (\text{D.6.16})$$

so that, since $\rho \geq \rho_n$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho^2}(f) \leq \mathcal{R}_{\rho(\rho+3\rho_n)}(f), \quad (\text{D.6.17})$$

Combining this bound with [Theorem D.1](#) (resp. [Theorem D.3](#)) completes the bound of [Theorem 6.3](#).

D.6.3 Regularized case

In the regularized case, the bound simplifies as well compared to [Proposition D.4.3](#).

Proposition D.6.5. *Fix a threshold $\delta \in (0, 1)$. When $\rho^2 \leq \rho_c^2(\varepsilon, \sigma)$, we have,*

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_{\rho^2}^\varepsilon(f) = \inf_{\underline{\lambda}_n \leq \lambda \leq \bar{\lambda}} \lambda \rho^2 + \mathbb{E}_{\xi \sim P} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \quad (\text{D.6.18})$$

where the dual bounds are defined by

$$\underline{\lambda} := \frac{\varepsilon}{\text{Var}(\varepsilon, \sigma)} \left(\rho_c^2(\varepsilon, \sigma) - \rho^2 \right) \quad (\text{D.6.19})$$

$$\text{and } \bar{\lambda} := \max \left(\frac{12\varepsilon}{R^2} \log(2 \times 6^{d/2}), e^{\frac{\sup_{f \in \mathcal{F}} \|f\|_\infty}{\varepsilon}} \frac{\varepsilon_0}{\rho} \right), \quad (\text{D.6.20})$$

and λ_0^*, μ^* were defined in [Assumption D.4](#).

Proof. The proof of the upper-bound is exactly the same as in [Proposition D.4.3](#) so we focus on the lower-bound. Following the same reasoning as the one to get (D.4.40) in [Proposition D.4.3](#) but with $\mathbb{P}$ instead of $\widehat{\mathbb{P}}_n$ we get that

$$\partial_\lambda \left\{ \lambda \rho^2 + \mathbb{E}_{\xi \sim \mathbb{P}} [\phi(f, \xi, \lambda, \varepsilon, \sigma)] \right\} \leq \rho^2 - \mathbb{E}_{\xi \sim \mathbb{P}} \left[\mathbb{E}_{\zeta \sim \pi_{\sigma}^{f/\varepsilon}(\cdot|\xi)} \left[\frac{1}{2} \|\xi - \zeta\|^2 \right] \right] + \frac{\lambda}{\varepsilon} \text{Var}(\varepsilon, \sigma) \quad (\text{D.6.21})$$

$$= \rho^2 - \rho_c^2(\varepsilon, \sigma) + \frac{\lambda}{\varepsilon} \text{Var}(\varepsilon, \sigma), \quad (\text{D.6.22})$$

which is non-positive when $\lambda \leq \underline{\lambda}$. $\blacksquare$

Corollary D.6.6. *In the same setting as [Theorem D.5](#), with probability $1 - \delta$, it holds,*

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_{\rho^2}^\varepsilon(f) \geq \widehat{\mathcal{R}}_{\rho(\rho - \rho_n)}^\varepsilon(f). \quad (\text{D.6.23})$$

Proof. This result follows by combining [Propositions D.3.1](#) and [D.6.1](#) and [Proposition D.6.5](#), which gives the desired result with probability at least $1 - \frac{\delta}{2}$ and *a fortiori* $1 - \delta$. $\blacksquare$

To conclude, in the context of [Theorem D.5](#), [Corollary D.6.6](#) with $\rho \leftarrow \rho + \rho_n$ and $\varepsilon_0 \leftarrow \frac{\varepsilon_0 \rho}{\rho + \rho_n}$, i.e., $\varepsilon \leftarrow \frac{\varepsilon_0 \rho}{\rho + \rho_n} \times (\rho + \rho_n)$, yields,² with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad \widehat{\mathcal{R}}_{\rho(\rho + \rho_n)}^{\varepsilon_0 \rho}(f) \leq \mathcal{R}_{(\rho + \rho_n)^2}^{\varepsilon_0 \rho}(f), \quad \text{and, in particular, } \widehat{\mathcal{R}}_{\rho^2}^{\varepsilon_0 \rho}(f) \leq \mathcal{R}_{\rho(\rho + 3\rho_n)}^{\varepsilon_0 \rho}(f). \quad (\text{D.6.24})$$

Combining this bound with [Theorem D.5](#) completes the bound of [Theorem 6.3](#).

D.7 TECHNICAL LEMMAS

In this section, we recall and adapt known results, as well as establish technical facts, all useful in our developments. They are presented in self-contained lemmas and are arranged in four thematic subsections.

D.7.1 Laplace approximation

Lemma D.7.1 (Restriction to Ξ). *Consider $\Xi \subset \mathbb{R}^d$, $\varepsilon_1, \tau_1 > 0$ and a map $\zeta^\star : [0, \tau_1] \rightarrow \Xi$ defined by $\zeta^\star(\tau) = \xi + \tau g$ with $\xi \in \Xi$, $g \in \mathbb{R}^d$ and assume that there is a positive radius R such that,*

1. *The closed ball $\overline{\mathbb{B}}(\zeta^\star(0), R)$ is included in Ξ .*
2. *R, τ_1 and $\|g\|$ satisfy $\frac{R^2}{6} \geq \tau_1^2 \|g\|^2$.*

² Though ε_0 now formally depends on ρ' , the same bounds still hold and do not become degenerate since ε_0 lies $[\varepsilon_0/2, \varepsilon_0]$ that avoids zero.

Then, for $(\varepsilon, \tau) \in [0, \varepsilon_1] \times [0, \tau_1]$,

$$\left| (2\pi\varepsilon\tau)^{-\frac{d}{2}} \int_{\Xi} \exp\left(-\frac{\|\zeta - \zeta^*(\tau)\|_2^2}{2\varepsilon\tau}\right) d\zeta - 1 \right| \leq 6^{d/2} e^{-\frac{R^2}{12\varepsilon\tau}}, \quad (\text{D.7.1})$$

Proof. The quantity to bound rewrites

$$\left| (2\pi\varepsilon\tau)^{-\frac{d}{2}} \int_{\Xi} \exp\left(-\frac{\|\zeta - \zeta^*(\tau)\|_2^2}{2\varepsilon\tau}\right) d\zeta - 1 \right| = (2\pi\varepsilon\tau)^{-\frac{d}{2}} \int_{\mathbb{R}^d \setminus \Xi} \exp\left(-\frac{\|\zeta - \zeta^*(\tau)\|_2^2}{2\varepsilon\tau}\right) d\zeta, \quad (\text{D.7.2})$$

so let us bound this integral. Since $\overline{\mathbf{B}}(\zeta^*(0), R)$ is inside Ξ , this means that, for any $\zeta \notin \Xi$, $\|\zeta - \xi\|$ is at least equal to R . Hence, for any $\zeta \notin \Xi$, one has that

$$\|\zeta - \zeta^*(\tau)\|^2 \geq \frac{1}{2} \|\zeta - \xi\|^2 - \tau^2 \|g\|^2 \quad (\text{D.7.3})$$

$$\geq \frac{1}{6} \|\zeta - \xi\|^2 + \frac{1}{6} R^2 + \frac{1}{6} R^2 - \tau^2 \|g\|^2 \quad (\text{D.7.4})$$

$$\geq \frac{1}{6} \|\zeta - \xi\|^2 + \frac{1}{6} R^2, \quad (\text{D.7.5})$$

so that we get the bound

$$(2\pi\varepsilon\tau)^{-\frac{d}{2}} \int_{\mathbb{R}^d \setminus \Xi} \exp\left(-\frac{\|\zeta - \zeta^*(\tau)\|_2^2}{2\varepsilon\tau}\right) d\zeta \leq e^{-\frac{R^2}{12\varepsilon\tau}} \times (2\pi\varepsilon\tau)^{-\frac{d}{2}} \int_{\mathbb{R}^d \setminus \Xi} \exp\left(-\frac{\|\zeta - \zeta\|_2^2}{12\varepsilon\tau}\right) d\zeta \quad (\text{D.7.6})$$

$$= 6^{d/2} e^{-\frac{R^2}{12\varepsilon\tau}}. \quad (\text{D.7.7})$$

■

D.7.2 Concentration

We rely on standard concentration tools that we encapsulate in the following lemma for convenience.

Lemma D.7.2. *Let $(\mathcal{X}, \text{dist})$ be a (totally bounded) separable metric space, $\mathbf{P}$ a probability distribution on a probability space Ξ and $\widehat{\mathbf{P}}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$ with $\xi_1, \dots, \xi_n \sim \mathbf{P}$ i.i.d.. Consider a mapping $X : \mathcal{X} \times \Xi \rightarrow \mathbb{R}$ and assume that,*

1. *For each $x \in \mathcal{X}$, $\xi \mapsto X(x, \xi)$ is measurable;*
2. *There is a constant $L > 0$ such that, for each $\xi \in \Xi$, $x \mapsto X(x, \xi)$ is L -Lipschitz;*
3. *X almost surely belongs to $[a, b]$.*

Then, for any $\delta \in (0, 1)$,

(a) *With probability at least $1 - \delta$,*

$$\forall x \in \mathcal{X}, \quad \mathbb{E}_{\xi \sim \widehat{\mathbf{P}}_n} [X(x, \xi)] - \mathbb{E}_{\xi \sim \mathbf{P}} [X(x, \xi)] \leq \frac{48L\mathcal{I}(\mathcal{X}, \text{dist})}{\sqrt{n}} + 2(b-a) \sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \quad (\text{D.7.8})$$

(b) With probability at least $1 - \delta$,

$$\forall x \in \mathcal{X}, \quad \mathbb{E}_{\xi \sim P}[X(x, \xi)] - \mathbb{E}_{\xi \sim \hat{P}_n}[X(x, \xi)] \leq \frac{48L\mathcal{I}(\mathcal{X}, \text{dist})}{\sqrt{n}} + 2(b-a)\sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \quad (\text{D.7.9})$$

Proof. First, let us note that we can assume that $\mathbb{E}_{\xi \sim P}[X(x, \xi)] = 0$ provided that we prove the bound above with the left-hand side divided by a factor two. Indeed, considering the random variables $Y(x, \xi) := X(x, \xi) - \mathbb{E}_{\xi \sim P}[X(x, \xi)]$, we see that Y satisfy the assumptions of the lemma, albeit with the constants $L \leftarrow 2L$, $a \leftarrow a - b$ and $b \leftarrow b - a$. Moreover, we only prove the assertion (a) since the (b) follows from (a) with $X \leftarrow -X$.

Step 1: Bound on the expectation. First, we focus on bounding the expectation of the quantity

$$\sup_{x \in \mathcal{X}} \{\mathbb{E}_{\xi \sim \hat{P}_n} X(x, \xi)\}. \quad (\text{D.7.10})$$

By the symmetrization principle (e.g., (Boucheron et al., Lem. 11.4)), with $s_1, \dots, s_n$ i.i.d. Rademacher random variables,

$$\mathbb{E} \left[\sup_{x \in \mathcal{X}} \{\mathbb{E}_{\xi \sim \hat{P}_n} X(x, \xi)\} \right] \leq 2\mathbb{E} \left[\sup_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n s_i X(x, \xi_i) \right]. \quad (\text{D.7.11})$$

Take $x, x' \in \mathcal{X}$. By the Lipschitz property of X , with $\xi \sim P$, for any $i = 1, \dots, n$, the random variable

$$\frac{s_i(X(x, \xi) - X(x', \xi))}{\sqrt{n}L} \quad (\text{D.7.12})$$

is bounded, in absolute value, by $\frac{\text{dist}(x, x')}{\sqrt{n}}$ and as such it is sub-Gaussian with parameter $\frac{\text{dist}(x, x')^2}{n}$ by Hoeffding's lemma (e.g., (Boucheron et al., Lem. 2.2)). As a consequence, by independence, the random variable

$$\sum_{i=1}^n \frac{s_i(X(x, \xi) - X(x', \xi))}{\sqrt{n}L} \quad (\text{D.7.13})$$

is sub-Gaussian with parameter $\text{dist}(x, x')^2$. Since, in addition, it is zero-mean, we can invoke Dudley's bound (e.g., (Boucheron et al., Cor. 13.2)) to get that,

$$\mathbb{E} \left[\sup_{x \in \mathcal{X}} \frac{1}{\sqrt{n}L} \sum_{i=1}^n s_i X(x, \xi_i) \right] \leq 12\mathcal{I}(\mathcal{X}, \text{dist}), \quad (\text{D.7.14})$$

or, in other words, by (D.7.12), that

$$\mathbb{E} \left[\sup_{x \in \mathcal{X}} \{\mathbb{E}_{\xi \sim \hat{P}_n} X(x, \xi)\} \right] \leq \frac{24L\mathcal{I}(\mathcal{X}, \text{dist})}{\sqrt{n}}. \quad (\text{D.7.15})$$

Step 2: Concentration inequality. Since the functions X are uniformly bounded, $\sup_{x \in \mathcal{X}} \{\mathbb{E}_{\xi \sim \hat{P}_n} X(x, \xi)\}$, seen as a function of $(\xi_1, \dots, \xi_n)$, satisfies the bounded difference property with constant $b - a$. Therefore, the bounded difference

inequality (e.g., (Boucheron et al., Thm. 6.2)) readily yields that, with probability at least $1 - \delta$,

$$\sup_{x \in \mathcal{X}} \left\{ \mathbb{E}_{\xi \sim \hat{\mathbb{P}}_n} X(x, \xi) \right\} \leq \mathbb{E} \sup_{x \in \mathcal{X}} \left\{ \mathbb{E}_{\xi \sim \hat{\mathbb{P}}_n} X(x, \xi) \right\} + (b - a) \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \quad (\text{D.7.16})$$

$$\leq \frac{24L\mathcal{I}(\mathcal{X}, \text{dist})}{\sqrt{n}} + (b - a) \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \quad (\text{D.7.17})$$

where we plugged in (D.7.15), the bound on the expectation from the first step. ■

D.7.3 Dudley's integral bounds

Lemma D.7.3. *Let $(\mathcal{X}_1, \text{dist}_1)$ and $(\mathcal{X}_2, \text{dist}_2)$ be two metric spaces, and consider $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ equipped with the distance $\text{dist} := c_1 \text{dist}_1 + c_2 \text{dist}_2$ with $c_1, c_2 > 0$. Then*

$$\mathcal{I}(\mathcal{X}, \text{dist}) \leq c_1 \mathcal{I}(\mathcal{X}_1, \text{dist}_1) + c_2 \mathcal{I}(\mathcal{X}_2, \text{dist}_2). \quad (\text{D.7.18})$$

Proof. Note that, for any $t > 0$, the inequality $N(t, \mathcal{X}, \text{dist}) \leq N(t, \mathcal{X}_1, c_1 \text{dist}_1) \times N(t, \mathcal{X}_2, c_2 \text{dist}_2)$ holds, so that, by subadditivity of the square root,

$$\mathcal{I}(\mathcal{X}, \text{dist}) = \int_0^{+\infty} \sqrt{\log N(t, \mathcal{X}, \text{dist})} dt \quad (\text{D.7.19})$$

$$\leq \int_0^{+\infty} \sqrt{\log N(t, \mathcal{X}_1, c_1 \text{dist}_1)} dt + \int_0^{+\infty} \sqrt{\log N(t, \mathcal{X}_2, c_2 \text{dist}_2)} dt \quad (\text{D.7.20})$$

$$= \int_0^{+\infty} \sqrt{\log N(t/c_1, \mathcal{X}_1, \text{dist}_1)} dt + \int_0^{+\infty} \sqrt{\log N(t/c_2, \mathcal{X}_2, \text{dist}_2)} dt \quad (\text{D.7.21})$$

$$= c_1 \mathcal{I}(\mathcal{X}_1, \text{dist}_1) + c_2 \mathcal{I}(\mathcal{X}_2, \text{dist}_2), \quad (\text{D.7.22})$$

where we performed changes of variable in to obtain the last equality. ■

Lemma D.7.4. *For $c > 0$,*

$$\mathcal{I}([0, c], |\cdot|) \leq \frac{c}{2} (1 + 2 \log 2). \quad (\text{D.7.23})$$

Proof. Noticing that $N(t, [0, c], |\cdot|) = 1$ whenever $t \geq c$, we get that

$$\mathcal{I}([0, c], |\cdot|) = \int_0^c \sqrt{\log N(t, [0, c], |\cdot|)} dt \quad (\text{D.7.24})$$

$$\leq \int_0^c (1 + \log N(t, [0, c], |\cdot|)) dt. \quad (\text{D.7.25})$$

Now, a rough bound on $N(t, [0, c], |\cdot|)$ is $1 + \frac{c}{t}$ which fits our purpose and yields

$$\mathcal{I}([0, c], |\cdot|) \leq c + \int_0^c \log \left(1 + \frac{c}{t} \right) dt = c(1 + 2 \log 2). \quad (\text{D.7.26})$$

■

D.7.4 Auxiliary results

We conclude these sections with auxiliary technical results.

The following lemma recalls basic inequalities with the logarithm function.

Lemma D.7.5. For $0 \leq x \leq \frac{1}{2}$, the following inequalities hold,

$$\log(1-x) \geq -2x \quad \text{and} \quad \log(1+x) \leq x. \quad (\text{D.7.27})$$

Lemma D.7.6. For $\alpha > 0$, the function $x \mapsto \frac{\log(\alpha+x)}{x}$ is non-increasing on $([e^{W(1)} - \alpha]_+, +\infty)$.

Proof. Denote by $f : x \mapsto \frac{\log(\alpha+x)}{x}$ this function, defined on $(0, +\infty)$. Its derivative is $f' : x \mapsto \frac{1}{x} \left(\frac{1}{x+\alpha} - \log(x+\alpha) \right)$. But the function $x \mapsto \frac{1}{x+\alpha} - \log(x+\alpha)$ is non-increasing, goes to $-\infty$ at infinity and its only potential zero is $e^{W(1)} - \alpha$ if it is positive,³ which yields the result. ■

Lemma D.7.7. For $\Xi \subset \mathbb{R}^d$ a compact set, $g \in \mathcal{C}(\Xi)$ and, $Q \in \mathcal{P}(\Xi)$,

$$\log \mathbb{E}_{\xi \sim Q} [e^{g(\xi)}] \leq \frac{\mathbb{E}_{\xi \sim Q} [g(\xi) e^{g(\xi)}]}{\mathbb{E}_{\xi \sim Q} [e^{g(\xi)}]}. \quad (\text{D.7.28})$$

Proof. Define $\phi : t \mapsto \log \mathbb{E}_{\xi \sim Q} [e^{t g(\xi)}]$ which is convex and differentiable, since g is continuous on the compact set Ξ . Hence,

$$0 = \phi(0) \geq \phi(1) + \phi'(1)(0-1), \quad (\text{D.7.29})$$

so that $\phi'(1) \geq \phi(1)$ which is the desired inequality. ■

Lemma D.7.8. For $a, b, c, r > 0$ fixed, consider the function defined on $\mathbb{R}_+$ by

$$\phi(\lambda) = a\lambda + \frac{b}{\lambda+r} - c \log(\lambda+r). \quad (\text{D.7.30})$$

Then, for any $\bar{\lambda} > 0$, ϕ is strongly convex on $[0, \bar{\lambda}]$ with strong convexity constant

$$\mu^* := \frac{2b}{(\bar{\lambda}+r)^3} + \frac{c}{(\bar{\lambda}+r)^2}. \quad (\text{D.7.31})$$

and the unique solution to the minimization problem

$$\min_{\lambda \geq 0} \phi(\lambda) \quad (\text{D.7.32})$$

is given by,

$$\lambda^* = \left[\frac{c + \sqrt{c^2 + 4ab}}{2a} - r \right]_+. \quad (\text{D.7.33})$$

Proof. ϕ is twice differentiable and its derivatives are, for $\lambda \geq 0$,

$$\phi'(\lambda) = a - \frac{b}{(\lambda+r)^2} - \frac{c}{\lambda+r} \quad (\text{D.7.34})$$

³ W denotes the Lambert function, i.e., the inverse of the map $x \mapsto xe^x$.

$$\phi''(\lambda) = \frac{2b}{(\lambda + r)^3} + \frac{c}{(\lambda + r)^2}, \quad (\text{D.7.35})$$

which shows that ϕ is strictly convex on $\mathbb{R}_+$ and yields its strong convexity on compact intervals. Then, the first order optimality condition $\phi'(\lambda) = 0$ gives us that

$$a(\lambda + r)^2 - c(\lambda + r) - b = 0, \quad (\text{D.7.36})$$

which has an unique solution satisfying $\lambda + r \geq 0$ which is given by,

$$\lambda^* = \frac{c + \sqrt{c^2 + 4ab}}{2a} - r. \quad (\text{D.7.37})$$

If $\lambda^* \geq 0$, then this is the solution we are looking for. If is not, this means that both roots of (D.7.36) are non-positive and therefore $\phi'(0) \geq 0$ which means that 0 is the solution to the minimization problem. ■

Lemma D.7.9. For $c > 0$, $\bar{\rho} > 0$ such that $\bar{\rho} \geq (4c)^{1/3}$, the inequality

$$\bar{\rho}^2 - \rho^2 - \frac{c}{\rho} \geq 0 \quad (\text{D.7.38})$$

holds in particular when

$$\frac{2c}{\bar{\rho}^2} \leq \rho \leq \bar{\rho} - \frac{2c}{\bar{\rho}^2} \quad (\text{D.7.39})$$

Proof. When $0 < \rho \leq \bar{\rho}$, the inequation $\bar{\rho}^2 - \rho^2 - \frac{c}{\rho} \geq 0$ is implied by

$$\bar{\rho}^2 \rho - \bar{\rho} \rho^2 - c \geq 0. \quad (\text{D.7.40})$$

Solving the latter yields the interval

$$\left[\frac{\bar{\rho}^2 - \sqrt{\bar{\rho}(\bar{\rho}^3 - 4c)}}{2\bar{\rho}}, \frac{\bar{\rho}^2 + \sqrt{\bar{\rho}(\bar{\rho}^3 - 4c)}}{2\bar{\rho}} \right] \quad (\text{D.7.41})$$

and the inequality $1 - u \leq \sqrt{1 - u}$ for $u \in [0, 1]$ yields the result. ■

THE RATE OF CONVERGENCE OF BREGMAN PROXIMAL
METHODS: LOCAL GEOMETRY VS. REGULARITY
VS. SHARPNESS

This appendix contains the proofs for [Chapter 7](#). It is extracted from [Azizian et al. \(2024d\)](#).

We provide here a series of basic properties, helper lemmas and auxiliary results that we use repeatedly in our paper.

E.1 LEMMAS ON NUMERICAL SEQUENCES

The first two results concern numerical sequences.

Lemma E.1.1. *Consider two sequences of real numbers $u_t, \alpha_t \geq 0, t = 1, 2, \dots$, such that*

$$u_{t+1} \leq u_t - \alpha_t u_t^{1+r} \quad \text{for some } r > 0 \text{ and all } t = 1, 2, \dots \quad (\text{E.1.1})$$

Then, for all $t = 1, 2, \dots$, we have:

$$u_{t+1} \leq \frac{u_1}{\left(1 + r u_1^r \sum_{s=1}^t \alpha_s\right)^{1/r}}. \quad (\text{E.1.2})$$

Proof. See ([Polyak, 1987](#), p. 46, Lem. 6). ■

The second result that we prove here is a slight variant of the above lemma. It is [Lemma E.1.2](#) that we restate for convenience.

Lemma E.1.2. *Suppose that $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ admits the asymptotic expansion*

$$f(x) = x - \lambda x^{1+r} + o(x^{1+r}) \quad \text{as } x \rightarrow 0 \quad (\text{E.1.3})$$

for positive constants $\lambda, r > 0$. Then, for $u_1 > 0$ small enough, the sequence $u_{t+1} = f(u_t), t = 1, 2, \dots$, converges to 0 at a rate of $u_t \sim (\lambda r t)^{-1/r}$.

Proof. By the assumption on f , there exists some $\varepsilon > 0$ such that

$$x - 2\lambda x^{1+r} \leq f(x) \leq x - \frac{\lambda}{2} x^{1+r} \quad \text{for all } x \in [0, \varepsilon]. \quad (\text{E.1.4})$$

Note first that, if $u_1 \leq \varepsilon$, [Lemma E.1.1](#) readily implies that u_t converges to 0 and that $u_t \leq \varepsilon$ for all t . Moreover, if ε is small enough so that $1 - 2\lambda \varepsilon^r > 0$ and u_1 is positive, this implies that all u_t , for $t = 1, 2, \dots$, are positive. In particular, we consider the sequence $u_t^{-r}, t = 1, 2, \dots$, for which we get

$$u_{t+1}^{-r} - u_t^{-r} = \left[u_t - \lambda u_t^{1+r} + o(u_t^{1+r}) \right]^{-r} - u_t^{-r}$$

$$= u_t^{-r}(1 - \lambda u_t^r + o(u_t^r))^{-r} - u_t^{-r} = r\lambda + o(1). \quad (\text{E.1.5})$$

Hence, $u_t^{-r} \sim r\lambda t$ which gives the result. $\blacksquare$

E.2 PROPERTIES OF BREGMAN DIVERGENCES AND THE INDUCED PROX-MAPPINGS

We recall some basic properties of the Bregman divergence and the induced prox-mapping. Variants of these properties are fairly well known in the literature, so we omit their proofs and we refer to [Beck and Teboulle \(2003\)](#), [Juditsky et al. \(2011\)](#), [Mertikopoulos and Zhou \(2019\)](#), [Mertikopoulos et al. \(2019\)](#). In particular, [Lemmas E.2.1–E.2.3](#) correspond to Lemmas B.1, B.2 and B.4(a) of [Mertikopoulos et al. \(2019\)](#), respectively.

Lemma E.2.1. *Let h be a Bregman regularizer on $\mathcal{X}$ and let ∇h be a continuous selection of ∂h . Then, for all $x \in \mathcal{X}_h$, $x^+ \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have:*

$$a) \quad \partial h(x) = \nabla h(x) + \text{PC}(x) \quad (\text{E.2.1a})$$

$$b) \quad x^+ = P_x(y) \iff \nabla h(x) + y \in \partial h(x^+) \iff \nabla h(x^+) - \nabla h(x) \in y - \text{PC}(x^+) \quad (\text{E.2.1b})$$

where $\text{PC}(x) = \{y \in \mathcal{Y} : \langle y, x' - x \rangle \leq 0 \text{ for all } x' \in \mathcal{X}\}$ denotes the polar cone to $\mathcal{X}$ at x .

Lemma E.2.2 (3-point identity). *For all $p \in \mathcal{X}$ and all $x, x^+ \in \mathcal{X}_h$, we have:*

$$D(p, x^+) = D(p, x) + D(x, x^+) + \langle \nabla h(x^+) - \nabla h(x), x - p \rangle \quad (\text{E.2.2})$$

Lemma E.2.3 (Non-expansiveness). *For all $x \in \mathcal{X}_h$ and all $y, y^+ \in \mathcal{Y}$ we have:*

$$\|P_x(y^+) - P_x(y)\| \leq \|y^+ - y\|_* \quad (\text{E.2.3})$$

The next two results that we provide consider the evolution of the Bregman divergence before and after a prox step (or two):

Lemma E.2.4. *Let $x^+ = P_x(y)$ for $x \in \mathcal{X}_h$, $y \in \mathcal{Y}$ such that x^+ is still in $\mathcal{U}'$. Then, for all $p \in \mathcal{X}$, $v \in \text{PC}(p)$, we have:*

$$D(p, x^+) \leq D(p, x) + \langle y - v, x^+ - p \rangle - D(x^+, x) \quad (\text{E.2.4a})$$

$$\leq D(p, x) + \langle y - v, x - p \rangle + \frac{1}{2}\|y - v\|_*^2. \quad (\text{E.2.4b})$$

Proof. Our proof follows ([Mertikopoulos et al., 2019](#), Proposition B.3), but with a slight modification to account for the extra term with $v \in \text{PC}(p)$. The first step is to invoke the three-point identity ([E.2.2](#)) to write

$$D(p, x) = D(p, x^+) + D(x^+, x) + \langle \nabla h(x) - \nabla h(x^+), x^+ - p \rangle. \quad (\text{E.2.5})$$

Then, after rearranging to isolate $D(p, x^+)$, we get

$$\begin{aligned} D(p, x^+) &= D(p, x) - D(x^+, x) - \langle \nabla h(x) - \nabla h(x^+), x^+ - p \rangle \\ &\leq D(p, x) - D(x^+, x) - \langle y, x^+ - p \rangle \end{aligned} \quad (\text{E.2.6})$$

where the inequality in the last line follows from [Lemma E.2.1](#). Hence, given that $\langle v, x^+ - p \rangle \leq 0$ by the fact that $v \in \text{PC}(p)$, we readily obtain

$$D(p, x^+) \leq D(p, x) - D(x^+, x) - \langle y - v, x^+ - p \rangle. \quad (\text{E.2.7})$$

For the second inequality of the lemma, note that

$$\begin{aligned} -\langle y - v, x^+ - p \rangle &= -\langle y - v, x^+ - x \rangle - \langle y - v, x - x^+ \rangle \\ &\leq \frac{1}{2} \|y - v\|_*^2 + \frac{1}{2} \|x^+ - x\|^2 - \langle y - v, x - x^+ \rangle \\ &\leq \frac{1}{2} \|y - v\|_*^2 + D(x^+, x) - \langle y - v, x - x^+ \rangle \end{aligned} \quad (\text{E.2.8})$$

where the penultimate inequality follows directly from Young's inequality and the last one from (7.18) since both x and x^+ are in $\mathcal{U}'$. Our assertion is then obtained by combining this last bound with (E.2.7). ■

Lemma E.2.5. *Let $x_i^+ = P_x(y_i)$ for some $x \in \mathcal{U}' \cap \mathcal{X}_h$ and $y_i \in \mathcal{Y}$ such that $x_i^+ \in \mathcal{U}'$, $i = 1, 2$. Then, for all $p \in \mathcal{X}$ and all $v \in \text{PC}(p)$, we have:*

$$D(p, x_2^+) \leq D(p, x) + \langle y_2 - v, x_1^+ - p \rangle + \frac{1}{2} \|y_2 - y_1 - v\|_*^2 - \frac{1}{2} \|x_1^+ - x\|^2. \quad (\text{E.2.9})$$

Proof. Our proof follows (Mertikopoulos et al., 2019, Proposition B.4), again with a slight modification to account for the extra terms with $v \in \text{PC}(p)$. Specifically, applying Lemma E.2.4 with $x_2^+ = P_x(y_2)$ and $v \in \text{PC}(p)$ gives

$$\begin{aligned} D(p, x_2^+) &\leq D(p, x) + \langle y_2 - v, x_2^+ - p \rangle - D(x_2^+, x) \\ &\leq D(p, x) + \langle y_2 - v, x_1^+ - p \rangle + \langle y_2 - v, x_2^+ - x_1^+ \rangle - D(x_2^+, x) \end{aligned} \quad (\text{E.2.10})$$

To lower bound $D(x_2^+, x)$, we use again Lemma E.2.4 with $p \leftarrow x_2^+$ and $x_1^+ = P_x(y_1)$, which gives

$$D(x_2^+, x_1^+) \leq D(x_2^+, x) + \langle y_1, x_1^+ - x_2^+ \rangle - D(x_1^+, x) \quad (\text{E.2.11})$$

and, after rearranging to isolate $D(x_2^+, x)$ and substituting the resulting bound in (E.2.10),

$$D(p, x_2^+) \leq D(p, x) + \langle y_2 - v, x_1^+ - p \rangle + \langle y_2 - y_1 - v, x_2^+ - x_1^+ \rangle - D(x_2^+, x_1^+) - D(x_1^+, x). \quad (\text{E.2.12})$$

Thus, by Young's inequality and the local strong convexity of h , we finally obtain

$$\begin{aligned} D(p, x_2^+) &\leq D(p, x) + \langle y_2 - v, x_1^+ - p \rangle + \frac{1}{2} \|y_2 - y_1 - v\|_*^2 \\ &\quad + \frac{1}{2} \|x_2^+ - x_1^+\|^2 - \frac{1}{2} \|x_2^+ - x_1^+\|^2 - \frac{1}{2} \|x_1^+ - x\|^2 \\ &\leq D(p, x) + \langle y_2 - v, x_1^+ - p \rangle + \frac{1}{2} \|y_2 - y_1 - v\|_*^2 - \frac{1}{2} \|x_1^+ - x\|^2 \end{aligned} \quad (\text{E.2.13})$$

and our proof is complete. ■

E.3 LEGENDRE EXPONENT FOR INTERIOR POINTS

We discussed in § 7.4.1 that $\beta_h(p) = 0$ whenever p is an interior point. We formalize this intuition below.

Lemma E.3.1. *Suppose that ∇h is locally Lipschitz continuous. Then $\beta_h(p) = 0$ for all $p \in \mathcal{X}_h$; in particular, $\beta_h(p) = 0$ whenever $p \in \text{ri } \mathcal{X}$.*

Proof. Fix some $p \in \mathcal{X}_h$ and suppose that ∇h is locally Lipschitz continuous. Then there exists a neighborhood $\mathcal{U}$ of p in $\mathcal{X}$ and some $K > 0$ such that

$$\|\nabla h(p) - \nabla h(x)\|_* \leq K \|p - x\| \quad \text{for all } x \in \mathcal{U} \cap \mathcal{X}_h. \quad (\text{E.3.1})$$

Now, since $\nabla h(p) \in \partial h(p)$, we also have

$$\begin{aligned} D(p, x) &= h(p) - h(x) - \langle \nabla h(x), p - x \rangle \leq \langle \nabla h(p) - \nabla h(x), p - x \rangle \\ &\leq \|\nabla h(p) - \nabla h(x)\|_* \|p - x\| \leq K \|p - x\|^2 \end{aligned} \quad (\text{E.3.2})$$

for all $x \in \mathcal{U} \cap \mathcal{X}_h$. This shows that (7.20) holds with $\beta = 0$, i.e., $\beta_h(p) = 0$. ■

E.4 A SEPARATION RESULT

We now prove Lemma E.4.1, which we restate below for convenience:

Lemma E.4.1. *Let $\mathcal{X}$ be a polyhedral domain of the form (7.60). Then, for all $x^* \in \mathcal{X}$, there exists $P = P(A, b, x^*) \geq 1$ such that, for all $\mathcal{I} \subseteq \mathcal{A} \equiv \mathcal{A}(x^*)$, at least one of the following holds:*

- (a) $\mathcal{I} \neq \emptyset$ and there exists $i \in \mathcal{A} \setminus \mathcal{I}$ such that $x_i \leq P \max\{x_j : j \in \mathcal{I}\}$ for all $x \in \mathcal{X}$.
- (b) There exists $z \in \ker A$ such that $\|z\| \leq P$, $z_i = 0$ if $i \in \mathcal{I}$ and $P \geq z_i \geq 1$ if $i \in \mathcal{A} \setminus \mathcal{I}$.

Proof. Our claim is trivial if $\mathcal{I} = \mathcal{A}$, so we will focus exclusively on the case $\mathcal{I} \subsetneq \mathcal{A}$. The stated constant $P = P(A, b, x^*)$ will then be obtained as the maximum of 1 and the constants we obtain for each possible $\mathcal{I} \subsetneq \mathcal{A}$.

The proof consists in discussing whether there exists $(\lambda_i)_{i \in \mathcal{A} \setminus \mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A} \setminus \mathcal{I}}$ not all zero and $(\mu_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$ such that the inclusion

$$\mathcal{X} \subset \left\{ x \in \mathbb{R}^n : \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \lambda_i x_i = \sum_{i \in \mathcal{I}} \mu_i x_i \right\} \quad (\text{E.4.1})$$

holds. Case (a) considers the case when such coefficients exist, while Case (b) when this is not possible.

CASE (A) Assume that there exists $(\lambda_i)_{i \in \mathcal{A} \setminus \mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A} \setminus \mathcal{I}}$ not all zero and $(\mu_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$ such that (E.4.1) holds. In this case, $\mathcal{I}$ must be non-empty since otherwise $\mathcal{X}$ would be reduced to $\{0\}$ (see the first inclusion), violating the definition (7.60) of $\mathcal{X}$. In addition, there is some $i \in \mathcal{A} \setminus \mathcal{I}$ such that $\lambda_i > 0$ and thus we have

$$\forall x \in \mathcal{X}, x_i \leq \frac{\max(|\lambda_j| : j \in \mathcal{I})}{\lambda_i} \max(x_j : j \in \mathcal{I}) \quad (\text{E.4.2})$$

which corresponds to the first case of the lemma.

CASE (B) For all $(\lambda_i)_{i \in \mathcal{A} \setminus \mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A} \setminus \mathcal{I}}$ not all zero and $(\mu_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$, (E.4.1) does not hold. To interpret this situation, we use the fact that $\mathcal{X}$ is of the general polyhedral form (7.60) so $\text{aff } \mathcal{X} = x^* + \ker A$ and x^* always satisfies $\sum_{i \in \mathcal{A} \setminus \mathcal{I}} \lambda_i x_i^* = \sum_{i \in \mathcal{I}} \mu_i x_i^* = 0$ so that

$$(\text{E.4.1}) \iff \text{aff } \mathcal{X} \subset \left\{ x \in \mathbb{R}^n : \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \lambda_i x_i = \sum_{i \in \mathcal{I}} \mu_i x_i \right\} \iff \ker A \subset \left\{ x \in \mathbb{R}^n : \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \lambda_i x_i = \sum_{i \in \mathcal{I}} \mu_i x_i \right\}$$

$$\iff \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \lambda_i e_i - \sum_{i \in \mathcal{I}} \mu_i e_i \in \text{row}(A).$$

Therefore, the fact that (E.4.1) does not hold for all $(\lambda_i)_{i \in \mathcal{A} \setminus \mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A} \setminus \mathcal{I}}$ not all zero and $(\mu_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$ means that, the system,

$$\sum_{i \in \mathcal{A} \setminus \mathcal{I}} \lambda_i e_i = \sum_{i \in \mathcal{I}} \mu_i e_i + A^\top r, \quad (\text{E.4.3})$$

with variables $(\lambda_i)_{i \in \mathcal{A} \setminus \mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A} \setminus \mathcal{I}}$ not all zero, $(\mu_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$, $r \in \mathbb{R}^m$, has no solution. Hence, by Motzkin's theorem on the alternative (see e.g., (Ben-Tal and Nemirovski, 2001, §1.4.2))¹, this means that the system,

$$\begin{cases} z_i > 0 & \text{for } i \in \mathcal{A} \setminus \mathcal{I} \\ z_i = 0 & \text{for } i \in \mathcal{A} \\ Az = 0 \end{cases} \quad (\text{E.4.4})$$

admits a solution $z \in \mathbb{R}^n$. Rescaling z and setting P to $\max(\|z\|, \|z\|_\infty)$ then gives the second case. ■

E.5 OMITTED CALCULATIONS

In this appendix, we provide some computational details that were left out of the main text to streamline our presentation.

Example 7.4 (Hellinger distance, [continuing](#) from p. 119). We proceed to compute the Taylor expansion of F near $x^* = -1$ for the shifted operator $g(x) = x + 1$. Indeed, in this case, the fixed point operator F is given by

$$\begin{aligned} F(x) &= P_x(-\gamma g(x)) = P_x(-\gamma(x+1)) \\ &= \frac{x - \gamma(x+1)\sqrt{1-x^2}}{\sqrt{1-x^2 + (x - \gamma(x+1)\sqrt{1-x^2})^2}} \\ &= \frac{G(x)}{\sqrt{1-x^2 + G(x)^2}}, \end{aligned} \quad (\text{E.5.1})$$

with $G(x) = x - \gamma(x+1)\sqrt{1-x^2}$. Now, the behavior of G near $x^* = -1$ can be approximated as

$$\begin{aligned} G(x) &= x - \gamma(x+1)^{3/2}(1-x)^{1/2} \\ &= -1 + (x+1) - \gamma(x+1)^{3/2}(2 - (x+1))^{1/2} \\ &= -1 + (x+1) - \sqrt{2}\gamma(x+1)^{3/2} \left(1 - \frac{1}{4}(x+1) + o(x+1)\right) \\ &= -1 + (x+1) - \sqrt{2}\gamma(x+1)^{3/2} + \frac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\left((x+1)^{5/2}\right). \end{aligned} \quad (\text{E.5.2})$$

Another Taylor expansion then yields

$$G(x)^2 = \left(1 - (x+1) + \sqrt{2}\gamma(x+1)^{3/2} - \frac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\left((x+1)^{5/2}\right)\right)^2$$

¹ With the notations of (Ben-Tal and Nemirovski, 2001, §1.4.2), the lines of the matrix S are made of the e_i for $i \in \mathcal{A} \setminus \mathcal{I}$ and the lines of the matrix N are the e_i for $i \in \mathcal{I}$, $-e_i$ for $i \in \mathcal{I}$, the lines of A and their opposite.

$$= 1 - 2(x+1) + 2\sqrt{2}\gamma(x+1)^{3/2} + (x+1)^2 - \frac{\sqrt{2}\gamma}{2}(x+1)^{5/2} + o\left((x+1)^{5/2}\right) \quad (\text{E.5.3})$$

so the denominator of (E.5.1) becomes

$$\begin{aligned} \sqrt{1 - x^2 + G(x)^2} &= \left((x+1)(2 - (x+1) + G(x)^2) \right)^{1/2} \\ &= \left(1 + 2\sqrt{2}\gamma(x+1)^{3/2} - \frac{\sqrt{2}\gamma}{2}(x+1)^{5/2} + o\left((x+1)^{5/2}\right) \right)^{1/2} \\ &= 1 + \sqrt{2}\gamma(x+1)^{3/2} - \frac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\left((x+1)^{5/2}\right). \end{aligned} \quad (\text{E.5.4})$$

Thus, plugging this expansion and (E.5.2) into (E.5.1) gives

$$\begin{aligned} F(x) &= \frac{-1 + (x+1) - \sqrt{2}\gamma(x+1)^{3/2} + \frac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\left((x+1)^{5/2}\right)}{1 + \sqrt{2}\gamma(x+1)^{3/2} - \frac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\left((x+1)^{5/2}\right)} \\ &= \left(-1 + (x+1) - \sqrt{2}\gamma(x+1)^{3/2} + \frac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\left((x+1)^{5/2}\right) \right) \\ &\quad \times \left(1 - \sqrt{2}\gamma(x+1)^{3/2} + \frac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\left((x+1)^{5/2}\right) \right) \\ &= -1 + (x+1) - 2\sqrt{2}\gamma(x+1)^{5/2} + o\left((x+1)^{5/2}\right), \end{aligned} \quad (\text{E.5.5})$$

which gives our assertion when $x^* = -1$. 

Example 7.8 (Three-dimensional simplex, [continuing](#) from p. 129). We conclude our treatment of the simplex by showing that $x_{2,t} \sim x_{2,t}/x_{3,t} = \Omega(1/t)$ if $v_2 = 0$ but $v_1 > 0$. To begin with, we have $g_2(x_t) = x_{2,t} = o(1)$ so, arguing as in the first part of the example, we readily get

$$\frac{x_{1,t+1}}{x_{2,t+1}} = \frac{x_{1,t}}{x_{2,t}} \exp(-\gamma v_1 + o(1)), \quad (\text{E.5.6})$$

so $x_{1,t}/x_{2,t}$ converges to 0 at a geometric rate. Accordingly, the quantity $x_{3,t}/x_{2,t}$ is bounded as

$$\begin{aligned} \frac{x_{3,t+1}}{x_{2,t+1}} &= \frac{x_{3,t}}{x_{2,t}} \exp(\gamma g_2(x_t) - \gamma g_3(x_t)) = \frac{x_{3,t}}{x_{2,t}} \exp(\gamma x_{2,t} - \gamma(x_{3,t} - 1)) \\ &= \frac{x_{3,t}}{x_{2,t}} \exp(2\gamma x_{2,t} + \gamma x_{1,t}) \leq \frac{x_{3,t}}{x_{2,t}} \exp\left(2\gamma \frac{x_{2,t}}{x_{3,t}} + \gamma x_{1,t}\right) \end{aligned} \quad (\text{E.5.7})$$

Now, since both $\frac{x_{2,t}}{x_{3,t}}$ and $x_{1,t}$ go to zero,

$$\begin{aligned} \frac{x_{3,t+1}}{x_{2,t+1}} &\leq \frac{x_{3,t}}{x_{2,t}} \left(1 + 2\gamma \frac{x_{2,t}}{x_{3,t}} + \gamma x_{1,t} + o\left(2\frac{x_{2,t}}{x_{3,t}} + x_{1,t}\right) \right) \\ &= \frac{x_{3,t}}{x_{2,t}} + 2\gamma + o(1). \end{aligned} \quad (\text{E.5.8})$$

since $x_{1,t}/x_{2,t}$ vanishes as $t \rightarrow \infty$. Hence, after telescoping, we conclude that $\frac{x_{3,t}}{x_{2,t}} \leq 2\gamma t + o(t)$, which in turn shows that $x_{2,t} \sim x_{2,t}/x_{3,t} = \Omega(1/t)$, as claimed. 

THE LAST-ITERATE CONVERGENCE RATE OF OPTIMISTIC
MIRROR DESCENT IN STOCHASTIC VARIATIONAL
INEQUALITIES

This appendix contains the proofs for § 7.6.1. It is extracted from [Azizian et al. \(2021b\)](#).

We gather in this section the basic results used in our proofs.

F.1 BREGMAN DIVERGENCES

We recall here classical lemmas on Bregman divergences; see [Juditsky et al. \(2011\)](#), [Mertikopoulos et al. \(2019\)](#). We complement them with two elementary results [Lemma F.1.6](#) on a specific bounding and [Lemma F.1.7](#) which shows that the Legendre exponent is typically trivial on $\mathcal{X}_h$.

Lemma F.1.1. *For $x \in \mathcal{X}$, $p \in \text{dom } \partial h$,*

$$D(x, p) \geq \frac{1}{2} \|p - x\|^2.$$

Lemma F.1.2. *For $x \in \text{dom } \partial h$, $y, y' \in \mathcal{Y}$,*

$$\|P_x(y) - P_x(y')\| \leq \|y - y'\|_*.$$

Lemma F.1.3. *For $p \in \mathcal{X}$, $x \in \text{dom } \partial h$, $y \in \mathcal{Y}$, $y' \in N_{\mathcal{X}}(p)$ and $x^+ = P_x(y)$,*

$$\begin{aligned} D(p, x^+) &\leq D(p, x) + \langle y - y', x^+ - p \rangle - D(x^+, x) \\ &\leq D(p, x) + \langle y - y', x - p \rangle + \frac{1}{2} \|y - y'\|_*^2. \end{aligned}$$

Lemma F.1.4. *For $p \in \mathcal{X}$, $x \in \text{dom } \partial h$, $y, y' \in \mathcal{Y}$ and $x_1^+ = P_x(y)$, $x_2^+ = P_x(y')$,*

$$D(p, x_2^+) \leq D(p, x) + \langle y', x_1^+ - p \rangle + \frac{1}{2} \|y' - y\|_*^2 - \frac{1}{2} \|x_1^+ - x\|^2.$$

Lemma F.1.5. *For $x \in \mathcal{X}_h$,*

$$\partial h(x) = \nabla h(x) + \text{NC}_{\mathcal{X}}(x).$$

As a consequence, $x \in \mathcal{X}_h$, $x^+ \in \mathcal{X}$, $y \in \mathcal{Y}$,

$$\begin{aligned} x^+ = P_x(y) &\iff \nabla h(x) + y \in \partial h(x^+) \\ &\iff \nabla h(x) + y - \nabla h(x^+) \in \text{NC}_{\mathcal{X}}(x^+). \end{aligned}$$

Moreover, $x^+ = P_x(y)$ implies that $x^+ \in \mathcal{X}_h$.

Lemma F.1.6. *If [Assumption 7.2](#) holds, $x^+ \in \mathcal{B}$, and $x \in \mathcal{X}$, then,*

$$\langle g(x^+), x^+ - x^* \rangle \geq \frac{\mu}{2} \|x - x^*\|^2 - \mu \|x^+ - x\|^2.$$

Proof. If $x^+ \in \mathcal{B}$, then,

$$\langle g(x^+), x^+ - x^* \rangle \geq \mu \|x^+ - x^*\|^2. \quad (\text{F.1.1})$$

However, we are interested in the distance between x and x^* , not in the distance between x^+ and x^* . To remedy this, we use Young's inequality,

$$\|x - x^*\|^2 \leq 2\|x - x^+\|^2 + 2\|x^+ - x^*\|^2,$$

which leads to,

$$\|x^+ - x^*\|^2 \geq \frac{1}{2} \|x - x^*\|^2 - \|x^+ - x\|^2.$$

Combined with (F.1.1), we get,

$$\langle g(x^+), x^+ - x^* \rangle \geq \frac{\mu}{2} \|x - x^*\|^2 - \mu \|x^+ - x\|^2.$$

■

Lemma F.1.7 (Trivial Legendre exponent on the interior). *Assume that ∇h is locally Lipschitz continuous on $\mathcal{X}_h$. Then, for any $p \in \mathcal{X}_h$, the Legendre exponent of h at p is $\beta = 0$.*

Proof. Take $p \in \mathcal{X}_h$. As ∇h is assumed to be locally Lipschitz, there exists $\mathcal{U}$ a neighborhood of p and $K > 0$ such that, for any $x \in \mathcal{U} \cap \mathcal{X}_h$,

$$\|\nabla h(p) - \nabla h(x)\|_* \leq \frac{K}{2} \|p - x\|.$$

Now, as $\nabla h(p) \in \partial h(p)$, for any $x \in \mathcal{U} \cap \mathcal{X}_h$,

$$\begin{aligned} D(p, x) &= h(p) - h(x) - \langle \nabla h(x), p - x \rangle \\ &\leq \langle \nabla h(p) - \nabla h(x), p - x \rangle \\ &\leq \|\nabla h(p) - \nabla h(x)\|_* \|p - x\|. \end{aligned}$$

Using the local Lipschitz continuity of ∇h , we get, for any $x \in \mathcal{U} \cap \mathcal{X}_h$,

$$D(p, x) \leq \frac{K}{2} \|p - x\|^2,$$

which reads $\beta = 0$. ■

F.2 SEQUENCES

We recall results on sequences to transform descent-like inequalities into rates; see the classical textbook [Polyak \(1987\)](#). We also need variants of existing results ([Lemma F.2.4](#) and [Lemma F.2.5](#)) that we state here.

Lemma F.2.1 ([Chung \(1954, Lem. 1\)](#)). *Let $(u_t)_{t \geq 1}$ be a sequence of non-negative scalars, $q > 1$, $q' > 0$, $t_0 \geq 0$. If, for any $t \geq 1$,*

$$a_{t+1} \leq \left(1 - \frac{q}{t + t_0}\right) a_t + \frac{q'}{(t + t_0)^2},$$

then, for any $T \geq 1$,

$$a_T \leq \frac{q'}{q-1} \frac{1}{T+t_0} + o\left(\frac{1}{T}\right).$$

Lemma F.2.2 (Chung (1954, Lem. 4)). Let $(a_t)_{t \geq 1}$ be a sequence of non-negative scalars, $q, q' > 0$, $t_0 \geq 0$, $1 > \eta > 0$. If, for any $t \geq 1$,

$$a_{t+1} \leq \left(1 - \frac{q}{(t+t_0)^\eta}\right) a_t + \frac{q'}{(t+t_0)^{2\eta}},$$

then, for any $T \geq 1$,

$$a_T \leq \mathcal{O}^* \frac{1}{T^\eta}.$$

Lemma F.2.3 (Polyak (1987, §2.2, Lem. 6)). Let $(a_t)_{1 \leq t \leq T}$ be a sequence of non-negative scalars, $(\gamma_t)_{1 \leq t \leq T}$ be sequence of positive scalars and $\alpha > 0$. If, for any $t = 1, \dots, T$,

$$a_{t+1} \leq a_t - \gamma_t a_t^{1+\alpha},$$

then,

$$a_T \leq \frac{a_1}{\left(1 + \alpha a_1^\alpha \sum_{t=1}^{T-1} \gamma_t\right)^{1/\alpha}}.$$

Though the next two lemmas allow for step-sizes of the form $\gamma_t = \frac{\gamma}{(t+t_0)^\eta}$, they do not exploit the term t_0 .

Lemma F.2.4. Let $(a_t)_{t \geq 1}$ be a sequence of non-negative scalars, $(\gamma_t)_{t \geq 1}$ be sequence of positive scalars of the form $\gamma_t = \frac{q}{(t+t_0)^\eta}$ with $q, \eta > 0$, $t_0 \geq 0$ and $\alpha > 0$, $q' > 0$ such that,

$$a_{t+1} \leq a_t - \gamma_t a_t^{1+\alpha} + \frac{q'}{(t+t_0)^{2\eta}}.$$

If,

$$1 \geq \eta \geq \frac{1+\alpha}{1+2\alpha} > \frac{1}{2}, \quad q' q^{1/\alpha} \leq c(\eta, \alpha) := \alpha \left(\frac{1-2^{-(1+\alpha)}}{1+\alpha} \right)^{1+\frac{1}{\alpha}},$$

then, for any $T \geq 1$,

$$a_T \leq \frac{a_1 + b}{\left(1 + \alpha(a_1 + b)^\alpha \sum_{t=1}^{T-1} \gamma_t\right)^{1/\alpha}},$$

where $b = \left(\frac{1-2^{1-2\eta}}{(1+\alpha)q} \right)^{\frac{1}{\alpha}}$.

Proof. First, define $p = 1 + \alpha > 1$ and note that $\frac{1+\alpha}{1+2\alpha} = \frac{p}{2p-1} = \frac{1}{2-p^{-1}} \in (\frac{1}{2}, 1)$ so, in particular, the condition on η is not absurd. Moreover, this means that $\beta := 2\eta - 1$ belongs to $(0, 1]$. Then, we have $b = \left(\frac{1-2^{-\beta}}{pq} \right)^{\frac{1}{\alpha}} > 0$ and, define, for $t \geq 1$, $b_t := \frac{b}{(t+t_0)^\beta}$.

The first part of the proof consists in showing, that, for $t \geq 1$,

$$\frac{q'}{(t+t_0)^{2\eta}} \leq b_t - b_{t+1} - \gamma_t b_t^p.$$

For this, we will need the following remark. As $\beta \leq 1$, $x \mapsto (1+x)^\beta$ is concave. Hence, it is above its chords, and in particular above its chord going from 0 to 1. Thus, for $0 \leq x \leq 1$, $(1+x)^\beta \geq 1+x(2^\beta-1)$.

We use this remark to lower bound $b_t - b_{t+1}$ for $t \geq 1$. Indeed,

$$\begin{aligned} b_t - b_{t+1} &= \frac{b}{(t+t_0)^\beta} - \frac{b}{(t+1+t_0)^\beta} \\ &= \frac{b}{(t+1+t_0)^\beta} \left(\left(1 + \frac{1}{t+t_0}\right)^\beta - 1 \right) \\ &\geq \frac{b}{(t+1+t_0)^\beta} \frac{2^\beta - 1}{t+t_0} \\ &\geq \frac{b}{(t+t_0)^{\beta+1}} \frac{2^\beta - 1}{2^\beta}. \end{aligned}$$

Therefore,

$$b_t - b_{t+1} - \gamma_t b_t^p \geq \frac{b}{(t+t_0)^{\beta+1}} (1 - 2^{-\beta}) - \frac{q b^p}{(t+t_0)^{\eta+p\beta}}.$$

Now, by the definition of β , $\beta + 1 = 2\eta$ and $\eta + p\beta = (2p+1)\eta - p = 2\eta + (2p-1)\eta - p \geq 2\eta$ by the assumption that $\eta \geq \frac{1+\alpha}{1+2\alpha} = \frac{p}{2p-1}$. Hence,

$$b_t - b_{t+1} - \gamma_t b_t^p \geq \frac{1}{(t+t_0)^{2\eta}} (b(1 - 2^{-\beta}) - q b^p),$$

so that we only need to show that $q' \leq b(1 - 2^{-\beta}) - q b^p$.

Rearranging and replacing b by its expression gives,

$$\begin{aligned} b(1 - 2^{-\beta}) - q b^p &= b((1 - 2^{-\beta}) - q b^\alpha) \\ &= b \left((1 - 2^{-\beta}) - \frac{1 - 2^{-\beta}}{p} \right) \\ &= b(1 - 2^{-\beta}) \frac{p-1}{p} \\ &= (1 - 2^{-\beta})^{1+\frac{1}{\alpha}} \frac{p-1}{p^{1+\frac{1}{\alpha}} q^{\frac{1}{\alpha}}}. \end{aligned}$$

Therefore, with $c(\eta, \alpha) = \alpha \left(\frac{1-2^{-(1+\alpha)}}{1+\alpha} \right)^{1+\frac{1}{\alpha}} = (1 - 2^{-\beta})^{1+\frac{1}{\alpha}} \frac{p-1}{p^{1+\frac{1}{\alpha}}} > 0$ and $q^{\frac{1}{\alpha}} q' \leq c(\eta, \alpha)$, we finally get that $q' \leq b(1 - 2^{-\beta}) - q b^p$ and, for $t \geq 1$,

$$\frac{q'}{(t+t_0)^{2\eta}} \leq b_t - b_{t+1} - \gamma_t b_t^p.$$

Recall that $(a_t)_{t \geq 1}$ satisfies, for $t \geq 1$,

$$a_{t+1} \leq a_t - \gamma_t a_t^p + \frac{q'}{(t+t_0)^{2\eta}}.$$

Therefore, putting these two inequalities together gives,

$$a_{t+1} + b_{t+1} \leq a_t + b_t - \gamma_t (a_t^p + b_t^p).$$

Finally, by convexity of $x \mapsto x^p$,

$$a_{t+1} + b_{t+1} \leq a_t + b_t - \gamma_t 2^{1-p} (a_t + b_t)^p.$$

Now, we can apply [Lemma F.2.3](#) with $a_t \leftarrow a_t + b_t$ and $\gamma_t \leftarrow \gamma_t 2^{1-p}$. For any $T \geq 1$,

$$a_T \leq a_T + b_T \leq \frac{1}{\left((a_1 + b_1)^{-\alpha} + \alpha 2^{-\alpha} \sum_{t=1}^{T-1} \gamma_t\right)^{1/\alpha}}.$$

As the right-hand side is non-decreasing in b_1 and $b_1 \leq b$, we get the result of the statement. $\blacksquare$

Lemma F.2.5. *Let $(a_t)_{t \geq 1}$ be a sequence of non-negative scalars, $(\gamma_t)_{t \geq 1}$ be sequence of positive scalars of the form $\gamma_t = \frac{q}{(t+t_0)^\eta}$ with $q, \eta > 0$, $b \geq 1$ and $\alpha > 0$, $q' > 0$ such that,*

$$a_{t+1} \leq a_t - \gamma_t a_t^{1+\alpha} + \frac{q'}{(t+t_0)^{2\eta}}.$$

If,

$$1 > \frac{1+\alpha}{1+2\alpha} \geq \eta > \frac{1}{2}, \quad q' q^{1/\alpha} \leq c(\eta, \alpha) := \frac{1-\eta}{(1+\alpha)^{\frac{1}{\alpha}+1}},$$

then, for any $T \geq 1$,

$$a_T \leq \frac{a_1}{\left(1 + \alpha a_1^\alpha \sum_{t=1}^{T-1} \gamma_t\right)^{1/\alpha}} + \frac{1}{((1+\alpha)q)^{\frac{1}{\alpha}}(T+t_0)^\eta}.$$

Proof. This proof is the “mirror” of the proof of the previous lemma. As before, define $p = 1 + \alpha > 1$ and note that $\frac{1+\alpha}{1+2\alpha} = \frac{p}{2p-1} = \frac{1}{2-p^{-1}} \in (\frac{1}{2}, 1)$ so the condition on η is not absurd. Moreover, this means that $\beta := \frac{\eta}{p}$ belongs to $(0, 1)$. Then, define $b := \left(\frac{1}{pq}\right)^{\frac{1}{\alpha}} > 0$ and, define, for $t \geq 1$, $b_t := \frac{b}{(t+t_0)^\beta}$.

Opposite to the proof of the previous lemma, the first part of the proof consists in showing, that, for $t \geq 1$,

$$\frac{q'}{(t+t_0)^{2\eta}} \leq b_{t+1} + \gamma_t b_t^p - b_t.$$

For this, we use the concavity of $x \mapsto (1+x)^\beta$, as $\beta \leq 1$, so that, for $x \geq 0$, $(1+x)^\beta \leq 1 + \beta x$.

This remark enables us to upper bound $b_t - b_{t+1}$ for $t \geq 1$. Indeed,

$$\begin{aligned} b_t - b_{t+1} &= \frac{b}{(t+t_0)^\beta} - \frac{b}{(t+1+t_0)^\beta} \\ &= \frac{b}{(t+1+t_0)^\beta} \left(\left(1 + \frac{1}{t+t_0}\right)^\beta - 1 \right) \\ &\leq \frac{b}{(t+1+t_0)^\beta} \frac{\beta}{t+t_0} \\ &\leq \frac{\beta b}{(t+t_0)^{\beta+1}}. \end{aligned}$$

Therefore,

$$b_t - b_{t+1} - \gamma_t b_t^p \leq \frac{\beta b}{(t + t_0)^{\beta+1}} - \frac{q b^p}{(t + t_0)^{\eta+p\beta}}.$$

Now, by the definition of $\beta = \frac{\eta}{p}$, $\eta + p\beta = 2\eta$ and $\beta + 1 = 2\eta + \frac{\eta}{p} + 1 - 2\eta = 2\eta + \frac{p-(2p-1)\eta}{p} \geq 2\eta$ by the assumption that $\eta \leq \frac{1+\alpha}{1+2\alpha} = \frac{p}{2p-1}$. Hence,

$$b_t - b_{t+1} - \gamma_t b_t^p \leq \frac{1}{(t + t_0)^{2\eta}} (\beta b - q b^p),$$

so that,

$$b_{t+1} + \gamma_t b_t^p - b_t \geq \frac{1}{(t + t_0)^{2\eta}} (q b^p - \beta b).$$

Again, we only need to show that $q' \leq q b^p - \beta b$.

Rearranging and replacing b by its expression gives,

$$\begin{aligned} q b^p - \beta b &= b(q b^{\alpha} - \beta) \\ &= b \left(\frac{1}{p} - \beta \right) \\ &= \frac{1}{q^{\frac{1}{\alpha}} p^{\frac{1}{\alpha}}} \left(\frac{1}{p} - \beta \right). \end{aligned}$$

Therefore, with $c(\eta, \alpha) = \frac{1}{p^{\frac{1}{\alpha}}} \left(\frac{1}{p} - \beta \right) > 0$ and $q^{\frac{1}{\alpha}} q' \leq c(\eta, \alpha)$, we finally get that $q' \leq q b^p - \beta b$ and, for $t \geq 1$,

$$\frac{q'}{(t + t_0)^{2\eta}} \leq b_{t+1} + \gamma_t b_t^p - b_t.$$

Putting this inequality together with the one on the sequence $(a_t)_{t \geq 1}$ gives,

$$a_{t+1} - b_{t+1} \leq a_t - b_t - \gamma_t (a_t^p - b_t^p).$$

Now, let us discuss separately the case when $a_t > b_t$ and when $a_t \leq b_t$.

More precisely, define $T_0 = \min\{t \geq 1 : a_t \leq b_t\} \in \mathbb{N}^* \cup \{+\infty\}$, so that, for any $1 \leq t < T_0$, $a_t > b_t > 0$. Note that, for $x, y > 0$, as $p > 1$, $x^p + y^p \leq (x + y)^p$. For $1 \leq t < T_0$, apply this with $x \leftarrow a_t - b_t$, $y \leftarrow b_t$ gives $(a_t - b_t)^p \leq a_t^p - b_t^p$ so that,

$$a_{t+1} - b_{t+1} \leq a_t - b_t - \gamma_t (a_t - b_t)^p.$$

Lemma F.2.3 with $a_t \leftarrow a_t - b_t$ gives, for $1 \leq T < T_0$,

$$\begin{aligned} a_T &\leq b_T + \frac{1}{((a_1 - b_1)^{-\alpha} + \alpha \sum_{t=1}^{T-1} \gamma_t)^{\frac{1}{\alpha}}} \\ &\leq b_T + \frac{1}{(a_1^{-\alpha} + \alpha \sum_{t=1}^{T-1} \gamma_t)^{\frac{1}{\alpha}}}, \end{aligned}$$

so the statement holds in this case.

We now handle the case of $T \geq T_0$. At $t = T_0$, we have $a_{T_0} \leq b_{T_0}$. We show, by induction, that, for any $t \geq T_0$, $a_t \leq b_t$. The initialization at $t = T_0$ is trivial. So, now, assume that $a_t \leq b_t$ for some $t \geq T_0$. Recall that we have,

$$a_{t+1} - b_{t+1} \leq a_t - b_t - \gamma_t (a_t^p - b_t^p),$$

so we only need to show that,

$$a_t - \gamma_t a_t^p \leq b_t - \gamma_t b_t^p.$$

Define $\varphi_t : x \in \mathbb{R}_+ \mapsto x - \gamma_t x^p$. Differentiating this function shows that it is non-decreasing on $\{x \in \mathbb{R}_+ : p\gamma_t x^{p-1} \leq 1\} = [0, (\gamma_t p)^{-\frac{1}{p-1}}]$. But $a_t \leq b_t \leq b_1 \leq b$ by definition of $(b_s)_{s \geq 1}$ and, by definition of b and $(\gamma_s)_{s \geq 1}$, $b = (pq)^{-\frac{1}{p-1}} = (\gamma_1 p)^{-\frac{1}{p-1}} \leq (\gamma_t p)^{-\frac{1}{p-1}}$.

Hence, both a_t and b_t belong to the interval on which φ_t is non-decreasing so that $\varphi_t(a_t) \leq \varphi_t(b_t)$, which implies that $a_{t+1} \leq b_{t+1}$. Therefore, we have shown that, for all $T \geq T_0$, $a_T \leq b_T = \frac{1}{(pq)^{\frac{1}{p-1}}(T+t_0)^{\frac{1}{p-1}}}$ which concludes the proof. ■

F.3 DESCENT INEQUALITY

The lemmas we present here link the divergence between two consecutive iterates and the solution.

In particular, the first one only rely on i) the definition of one iteration of optimistic mirror descent; and ii) the core properties of the Bregman regularizer i.e., the 1-strong convexity of h (through [Lemma F.1.1](#)) and the non-expansivity of the proximal mapping ([Lemma F.1.2](#)). It does not require any assumption on the vector field g , on the noise in g , or additional properties of the Bregman divergence (such as a Legendre exponent).

Lemma F.3.1. *If the stepsizes verify $\mu\gamma_t + 4L^2\gamma_t^2 \leq 1$, then, the iterates of optimistic mirror descent initialized with $x_{1/2} \in \mathcal{X}$, $x_1 \in \text{dom } \partial h$ verify for all $t \geq 1$,*

$$\begin{aligned} D(x^*, x_{t+1}) + \phi_{t+1} &\leq D(x^*, x_t) + (1 - \gamma_t \mu) \phi_t \\ &\quad - \gamma_t \langle g_{t+1/2}, x_{t+1/2} - x^* \rangle + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right) \|x_{t+1/2} - x_t\|^2 \\ &\quad + \gamma_t^2 \Delta_{t+1}^2 - \gamma_t^2 \tau^2 \|x_t - x_{t-1/2}\|^2, \end{aligned}$$

where $\Delta_{t+1}^2 = (\|g_{t+1/2} - g_{t-1/2}\|^2 - L^2 \|x_{t+1/2} - x_{t-1/2}\|^2)_+$, $\phi_t = \frac{\gamma_{t-1}^2}{2} \|g_{t-1/2} - g_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = \frac{1}{2} \|x_1 - x_{1/2}\|^2$.

Proof. First, apply [Lemma F.1.4](#) with $(x, y_1, y_2) \leftarrow (x_t, -\gamma_t g_{t-1/2}, -\gamma_t g_{t+1/2})$ and $p \leftarrow x^*$,

$$D(x^*, x_{t+1}) \leq D(x^*, x_t) - \gamma_t \langle g_{t+1/2}, x_{t+1/2} - x^* \rangle + \frac{\gamma_t^2}{2} \|g_{t-1/2} - g_{t+1/2}\|^2 - \frac{1}{2} \|x_{t+1/2} - x_t\|^2. \quad (\text{F.3.1})$$

In order to upper-bound $\frac{\gamma_t^2}{2} \|g_{t-1/2} - g_{t+1/2}\|^2 - \frac{1}{2} \|x_{t+1/2} - x_t\|^2$, we use the definition of $\phi_{t+1} = \frac{\gamma_t^2}{2} \|g_{t+1/2} - g_{t-1/2}\|^2$ to see that

$$\begin{aligned} \frac{\gamma_t^2}{2} \|g_{t-1/2} - g_{t+1/2}\|^2 &= \gamma_t^2 \|g_{t-1/2} - g_{t+1/2}\|^2 - \frac{\gamma_t^2}{2} \|g_{t-1/2} - g_{t+1/2}\|^2 \\ &= \gamma_t^2 \|g_{t-1/2} - g_{t+1/2}\|^2 - \phi_{t+1}, \end{aligned}$$

and the definition of Δ_{t+1} to bound

$$\begin{aligned} \gamma_t^2 \|g_{t-1/2} - g_{t+1/2}\|^2 &\leq \gamma_t^2 L^2 \|x_{t-1/2} - x_{t+1/2}\|^2 + \gamma_t^2 \Delta_{t+1}^2 \\ &\leq 2\gamma_t^2 L^2 \|x_{t-1/2} - x_t\|^2 + 2\gamma_t^2 L^2 \|x_{t+1/2} - x_t\|^2 + \gamma_t^2 \Delta_{t+1}^2, \end{aligned}$$

where we used Young's inequality to get the last inequality line. Putting the two equations together, we get that

$$\begin{aligned} & \frac{\gamma_t^2}{2} \|g_{t-1/2} - g_{t+1/2}\|^2 - \frac{1}{2} \|x_{t+1/2} - x_t\|^2 \\ & \leq 2\gamma_t^2 L^2 \|x_{t-1/2} - x_t\|^2 + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right) \|x_{t+1/2} - x_t\|^2 - \phi_{t+1} + \gamma_t^2 \Delta_{t+1}^2. \end{aligned} \quad (\text{F.3.2})$$

For $t \geq 2$, we use the non-expansivity of the proximal mapping (Lemma F.1.2) to bound $\|x_{t-1/2} - x_t\|$ by $\gamma_{t-1} \|g_{t-3/2} - g_{t-1/2}\|_*$ in (F.3.2) to get

$$\begin{aligned} & \frac{\gamma_t^2}{2} \|g_{t-1/2} - g_{t+1/2}\|^2 - \frac{1}{2} \|x_{t+1/2} - x_t\|^2 \\ & \leq (2\gamma_t^2 L^2) \gamma_{t-1}^2 \|g_{t-3/2} - g_{t-1/2}\|_*^2 + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right) \|x_{t+1/2} - x_t\|^2 - \phi_{t+1} + \gamma_t^2 \Delta_{t+1}^2 \\ & = (4\gamma_t^2 L^2) \phi_t + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right) \|x_{t+1/2} - x_t\|^2 - \phi_{t+1} + \gamma_t^2 \Delta_{t+1}^2 \\ & \leq (1 - \mu\gamma_t) \phi_t + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right) \|x_{t+1/2} - x_t\|^2 - \phi_{t+1} + \gamma_t^2 \Delta_{t+1}^2, \end{aligned}$$

where we used the definition of ϕ_t and the assumption on the stepsizes. Combining this result with (F.3.1) gives the first assertion of the lemma for $t \geq 2$.

Now, if $t = 1$, (F.3.2) can be rewritten as,

$$\begin{aligned} & \frac{\gamma_1^2}{2} \|g_{1/2} - g_{3/2}\|^2 - \frac{1}{2} \|x_{3/2} - x_1\|^2 \\ & \leq 2\gamma_1^2 L^2 \|x_{1/2} - x_1\|^2 + \left(2\gamma_1^2 L^2 - \frac{1}{2}\right) \|x_{3/2} - x_1\|^2 - \phi_2 + \gamma_1^2 \Delta_2^2 \\ & = 4\gamma_1^2 L^2 \phi_1 + \left(2\gamma_1^2 L^2 - \frac{1}{2}\right) \|x_{3/2} - x_1\|^2 - \phi_2 + \gamma_1^2 \Delta_2^2, \end{aligned}$$

which yields the assertion of the lemma for $t = 1$ as $4\gamma_1^2 L^2 \leq 1 - \gamma_1 \mu$. $\blacksquare$

We now use the properties of g and g to refine the bound of Lemma F.3.1 into a descent inequality.

Lemma F.3.2. *Let Assumption 7.1 and Assumption 7.2 hold. Consider the iterates of optimistic mirror descent with:*

i) *an unbiased stochastic oracle $g_t = g(x_t) + U_t$ (see (7.93));*

ii) *step-sizes $0 < \gamma_t \leq \frac{1}{4L}$.*

Then,

$$\begin{aligned} D(x^*, x_{t+1}) + \phi_{t+1} & \leq D(x^*, x_t) + (1 - \gamma_t \mu) \phi_t - \gamma_t \langle g_{t+1/2}, x_{t+1/2} - x^* \rangle \\ & \quad + \left(4\gamma_t^2 L^2 - \frac{1}{2}\right) \|x_{t+1/2} - x_t\|^2 + 4\gamma_t^2 (\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2) \end{aligned} \quad (\text{F.3.3})$$

where $\phi_t = \frac{\gamma_{t-1}^2}{2} \|g_{t-1/2} - g_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = \frac{1}{2} \|x_1 - x_{1/2}\|^2$.

Proof. First, let us examine the choice of γ_t . The condition $\gamma_t \leq \frac{1}{4L}$ is actually equivalent to,

$$8\gamma_t^2 L^2 + 2\gamma_t L \leq 1.$$

As $\mu \leq L$, this implies that,

$$8\gamma_t^2 L^2 + 2\gamma_t \mu \leq 1. \quad (\text{F.3.4})$$

We first use the inequality of OMD provided by [Lemma F.3.1](#) with $(\mu, L) \leftarrow (\mu, \sqrt{2}L)$ (note that its assumption is satisfied thanks to (F.3.4)) to get

$$\begin{aligned} D(x^*, x_{t+1}) + \phi_{t+1} &\leq D(x^*, x_t) + (1 - \gamma_t \mu) \phi_t \\ &\quad - \gamma_t \langle g_{t+1/2}, x_{t+1/2} - x^* \rangle + \left(2\gamma_t^2 L^2 - \frac{1}{2} \right) \|x_{t+1/2} - x_t\|^2 \\ &\quad + \gamma_t^2 \Delta_{t+1}^2 - \gamma_t^2 \tau^2 \|x_t - x_{t-1/2}\|^2, \end{aligned} \quad (\text{F.3.5})$$

where $\Delta_{t+1}^2 = (\|g_{t+1/2} - g_{t-1/2}\|_*^2 - L^2 \|x_{t+1/2} - x_{t-1/2}\|^2)_+$, $\phi_t = \frac{\gamma_{t-1}^2}{2} \|g_{t-1/2} - g_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = \frac{1}{2} \|x_1 - x_{1/2}\|^2$.

We now have to bound $\Delta_{t+1}^2 = (\|g_{t+1/2} - g_{t-1/2}\|_*^2 - 2L^2 \|x_{t+1/2} - x_{t-1/2}\|^2)_+$ for $t \geq 1$. By using twice Young's inequality,

$$\begin{aligned} \|g_{t+1/2} - g_{t-1/2}\|_*^2 &\leq 2\|g(x_{t+1/2}) - g(x_{t-1/2})\|_*^2 + 2\|U_{t+1/2} - U_{t-1/2}\|_*^2 \\ &\leq 2\|g(x_{t+1/2}) - g(x_{t-1/2})\|_*^2 + 4\|U_{t+1/2}\|_*^2 + 4\|U_{t-1/2}\|_*^2. \end{aligned}$$

Using the Lipschitz continuity of g ([Assumption 7.1](#)), we obtain that

$$\|g_{t+1/2} - g_{t-1/2}\|_*^2 \leq 2L^2 \|x_{t+1/2} - x_{t-1/2}\|^2 + 4\|U_{t+1/2}\|_*^2 + 4\|U_{t-1/2}\|_*^2,$$

so that,

$$\Delta_{t+1}^2 \leq 4\|U_{t+1/2}\|_*^2 + 4\|U_{t-1/2}\|_*^2.$$

Plugging this inequality in (F.3.5), we obtain the claimed result. $\blacksquare$

F.4 STOCHASTIC LOCAL STABILITY

In order to use the local properties of the Bregman divergence around the solution, we need a local stability result that gives us that sufficiently close iterates will remain in a neighborhood of the solution with high probability. To show such a result, we will use the following lemma [Hsieh et al. \(2020, Lem. F.1\)](#) for bounding a recursive stochastic process.

Lemma F.4.1 ([Hsieh et al. \(2020, Lem. F.1\)](#)). *Consider a filtration $(\mathcal{F}_t)_t$ and four $(\mathcal{F}_t)_t$ -adapted processes $(D_t)_t$, $(\zeta_t)_t$, $(\chi_t)_t$, $(\xi_t)_t$ such that χ_t is non-negative and the following recursive inequality is satisfied for all $t \geq 1$*

$$D_{t+1} \leq D_t - \zeta_t + \xi_t.$$

For $C > 0$, we define the events A_t by $A_1 := \{D_1 \leq C/2\}$ and $A_t := \{D_t \leq C\} \cap \{\chi_t \leq C/4\}$ for $t \geq 2$. We consider also the decreasing sequence of events I_t defined by $I_t := \bigcap_{1 \leq s \leq t} A_s$. If the following three assumptions hold true

$$(i) \quad \forall t, \zeta_t \mathbb{1}_{I_t} \geq 0,$$

$$(ii) \quad \forall t, \mathbb{E}[\xi_t | \mathcal{F}_t] \mathbb{1}_{I_t} = 0,$$

$$(iii) \quad \sum_{t=1}^{\infty} \mathbb{E}[(\xi_{t+1}^2 + \chi_{t+1}) \mathbb{1}_{I_t}] \leq \delta \epsilon \mathbb{P}(A_1),$$

where $\epsilon = \min(C^2/16, C/4)$ and $\delta \in (0, 1)$, then $\mathbb{P}(\bigcap_{t \geq 1} A_t | A_1) \geq 1 - \delta$.

Lemma F.4.2. *Let [Assumption 7.1](#) and [Assumption 7.2](#) hold. Consider the iterates of optimistic mirror descent with:*

i) an unbiased stochastic oracle $g_t = g(x_t) + U_t$ with finite variance σ^2 (see (7.93) and (7.94));

ii) step-sizes $0 < \gamma_t \leq \frac{1}{4L}$ with $\sum_{t=1}^{\infty} \gamma_t^2 = \Gamma < +\infty$.

Denote by $\mathcal{U}$ the neighborhood of x^* on which (7.19) holds with $p \leftarrow x^*$. Then, for any $r > 0$ such that $\mathcal{U} := \mathcal{B}(x^*, r) \cap \mathcal{X}$ is included in both $\mathcal{U}$ and $\mathcal{B}$,

$$\mathbb{P} \left[\forall t \geq \frac{1}{2}, x_t \in \mathcal{U} \mid D(x^*, x_1) + \phi_1 \leq \frac{2r^2}{9} \right] \geq 1 - \frac{9(8+r^2)\sigma^2\Gamma}{r^2 \min(1, r^2/9)},$$

where $\phi_1 = \frac{1}{2}\|x_1 - x_{1/2}\|^2$.

Proof. This proof mainly consists in applying Lemma F.4.1 indexed by $t \in \frac{1}{2}\mathbb{N}^*$, $t \geq 1$. Define $C = \frac{r^2}{2+1/4}$ and, for $t \geq 1$, the following adapted processes,

$$\begin{aligned} \zeta_t &:= 0 & \zeta_{t+1/2} &:= \gamma_t \langle g(x_{t+1/2}), x_{t+1/2} - x^* \rangle \\ \chi_{t+1/2} &:= 4\gamma_t^2 \|U_t\|_*^2 & \chi_{t+1} &:= 4\gamma_t^2 \|U_{t+1/2}\|_*^2 \\ \xi_{t+1/2} &:= 0 & \xi_{t+1} &:= -\gamma_t \langle U_{t+1/2}, x_{t+1/2} - x^* \rangle \\ D_t &:= D(x^*, x_t) + \phi_t & D_{t+1/2} &:= D_t - \zeta_t + \chi_{t+1/2} + \xi_{t+1/2}. \end{aligned}$$

With these definitions, for any $t \in \mathbb{N}^*$, $t \geq 1$,

$$\begin{aligned} D_{t+1/2} &\leq D_t - \zeta_t + \chi_{t+1/2} + \xi_{t+1/2} \\ D_{t+1} &\leq D_{t+1/2} - \zeta_{t+1/2} + \chi_{t+1} + \xi_{t+1} \end{aligned}$$

where the first inequality comes directly from the definition while the second one comes from the descent inequality of Lemma F.3.2.

Now define the events, for $t \in \frac{1}{2}\mathbb{N}^*$, $t \geq 1$,

$$I_t = \{D_1 \leq C/2\} \cap \bigcap_{s \in \frac{1}{2}\mathbb{N}: 3/2 \leq s \leq t} \{D_s \leq C\} \cap \{\chi_s^2 \leq C/4\}.$$

We first show, by induction on $t \in \frac{1}{2}\mathbb{N}^*$, $t \geq 1$, that, $I_t \subset \{x_s \in \mathcal{U}, s = \frac{1}{2}, 1, \dots, t - \frac{1}{2}, t\}$.

INITIALIZATION: For $t = 1$, the fact that $I_1 \subset \{x_{1/2} \in \mathcal{U}\}$ comes from the conditioning that $D(x^*, x_1) + \phi_1 \leq C/2$ and by definition of ϕ_1 and the strong convexity of h (Lemma F.1.1),

$$\|x_{1/2} - x^*\|^2 \leq 2(\|x_1 - x^*\|^2 + \|x_1 - x_{1/2}\|^2) \leq 4(D(x^*, x_1) + \phi_1) \leq 2C \leq r^2.$$

To show that $x_1 \in \mathcal{U}$, use the strong convexity of h (Lemma F.1.1) to get that, on I_1 ,

$$\|x_1 - x^*\|^2 \leq 2D(x^*, x_1) \leq 2D_1 \leq C \leq r^2.$$

INDUCTION STEP FOR ITERATES: Assume that, for some $t \geq 2$, $I_{t-1/2} \subset \{x_s \in \mathcal{U}, s = \frac{1}{2}, 1, \dots, t-1, t - \frac{1}{2}\}$. By definition, the sequence of events (I_s) is non-increasing so that, $I_t \subset \{x_s \in \mathcal{U}, s = \frac{1}{2}, 1, \dots, t-1, t - \frac{1}{2}\}$. Hence, we only have to show that $I_t \subset \{x_t \in \mathcal{U}\}$. But, by definition, $I_t \subset \{D_t \leq C\}$. Again, using the strong convexity of h (Lemma F.1.1), on I_t ,

$$\|x_t - x^*\|^2 \leq 2D(x^*, x_t) \leq 2D_t \leq 2C \leq r^2.$$

INDUCTION STEP FOR HALF-ITERATES: Assume that, for some $t \geq 1$, $I_t \subset \{x_s \in \mathcal{U}, s = \frac{1}{2}, 1, \dots, t-1, t\}$. By definition, the sequence of events (I_s) is

non-increasing so that, $I_{t+1/2} \subset \{x_s \in \mathcal{U}, s = \frac{1}{2}, 1, \dots, t-1, t\}$. So we focus on showing that $I_{t+1/2} \subset \{x_{t+1/2} \in \mathcal{U}\}$. For this, apply the first statement of [Lemma F.1.3](#) with $(x, p, y, y') \leftarrow (x_t, x^*, -\gamma_t(g(x_{t-1/2}) + U_t), -\gamma_t g(x^*))$,

$$D(x^*, x_{t+1/2}) \leq D(x^*, x_t) - \gamma_t \langle g(x_{t-1/2}) + U_t - g(x^*), x_{t+1/2} - x^* \rangle,$$

and apply Young's inequality twice to get

$$\begin{aligned} D(x^*, x_{t+1/2}) &\leq D(x^*, x_t) + \gamma_t^2 \|g(x_{t-1/2}) + U_t - g(x^*)\|_*^2 + \frac{1}{4} \|x_{t+1/2} - x^*\|^2 \\ &\leq D(x^*, x_t) + 2\gamma_t^2 \|g(x_{t-1/2}) - g(x^*)\|_*^2 + 2\gamma_t^2 \|U_t\|^2 + \frac{1}{4} \|x_{t+1/2} - x^*\|^2. \end{aligned}$$

By the strong convexity of h ([Lemma F.1.1](#)) and the Lipschitz continuity of g ([Assumption 7.1](#)),

$$\frac{1}{4} \|x_{t+1/2} - x^*\|^2 \leq D(x^*, x_t) + 2\gamma_t^2 L^2 \|x_{t-1/2} - x^*\|^2 + 2\gamma_t^2 \|U_t\|^2.$$

But, by definition, on $I_{t+1/2}$, $D(x^*, x_t) \leq C$, $x_{t-1/2}$ is in $\mathcal{U}$ and $\chi_{t+1/2} = 4\gamma_t^2 \|U_t\|_*^2 \leq C/4$. Therefore,

$$\frac{1}{4} \|x_{t+1/2} - x^*\|^2 \leq \left(1 + \frac{1}{8}\right) C + 2\gamma_t^2 L^2 r^2.$$

Using the definition of C and the bound on the step-size $\gamma_t \leq 1/(4L)$,

$$\|x_{t+1/2} - x^*\|^2 \leq \frac{r^2}{2} + \frac{r^2}{2} = r^2,$$

which concludes the induction step.

We now verify the assumptions needed to apply [Lemma F.4.1](#):

- (i) For $t \in \frac{1}{2}\mathbb{N}^*$, $t \geq 1$, $\zeta_t \mathbb{1}_{I_t} \geq 0$. If $t \in \mathbb{N}^*$, this is trivial as $\zeta_t = 0$. Now, fix $t \in \mathbb{N}^*$, $\zeta_{t+1/2} = \gamma_t \langle g(x_{t+1/2}), x_{t+1/2} - x^* \rangle$. But, on $I_{t+1/2}$, $x_{t+1/2} \in \mathcal{U}$ and so, by monotonicity of g ([Assumption 7.2](#)), $\langle g(x_{t+1/2}), x_{t+1/2} - x^* \rangle \geq 0$ on $I_{t+1/2}$.
- (ii) For $t \in \frac{1}{2}\mathbb{N}^*$, $t \geq 1$, $\mathbb{E} [\xi_{t+1/2} \mid \mathcal{F}_t] \mathbb{1}_{I_t} = 0$. There is only something to prove when t is of the form $t + \frac{1}{2}$ with $t \in \mathbb{N}^*$. In this case, as $x_{t+1/2}$ is $\mathcal{F}_{t+1/2}$ measurable,

$$\mathbb{E} [\xi_{t+1} \mid \mathcal{F}_{t+1/2}] = -\gamma_t \langle \mathbb{E} [U_{t+1/2} \mid \mathcal{F}_{t+1/2}], x_{t+1/2} - x^* \rangle = 0.$$

- (iii) $\sum_{t \in \frac{1}{2}\mathbb{N}, t \geq 1} \mathbb{E} \left(\left(\xi_{t+1/2}^2 + \chi_{t+1/2} \right) \mathbb{1}_{I_t} \right) \leq \delta \epsilon \mathbb{P}(I_1)$ with $\epsilon = \min(C^2/16, C/4) = \min(C/4, 1)C/4$ and $\delta = \frac{\sigma^2(8+r^2)\Gamma}{\epsilon}$. For this let us first bound each of the terms involved individually. For $t \in \mathbb{N}^*$, by assumption on the noise and as I_t is $\mathcal{F}_t$ measurable, and so $\mathcal{F}_{t+1/2}$ measurable,

$$\begin{aligned} \mathbb{E}(\chi_{t+1/2} \mathbb{1}_{I_t}) &= 4\gamma_t^2 \mathbb{E} \left(\|U_{t+1/2}\|_*^2 \mathbb{1}_{I_t} \right) \\ &= 4\gamma_t^2 \mathbb{E} \left(\mathbb{E} [\|U_{t+1/2}\|_*^2 \mid \mathcal{F}_{t+1/2}] \mathbb{1}_{I_t} \right) \\ &\leq 4\gamma_t^2 \sigma^2 \mathbb{P}(I_t) \\ &\leq 4\gamma_t^2 \sigma^2 \mathbb{P}(I_1), \end{aligned}$$

where we used that the sequence of events $(I_s)_s$ is non-increasing. Likewise, $\mathbb{E}(\chi_{t+1} \mathbb{1}_{I_{t+1/2}}) \leq 4\gamma_t^2 \sigma^2 \mathbb{P}(I_1)$. Now, by definition of the dual norm,

$$\begin{aligned} \mathbb{E}(\xi_{t+1}^2 \mathbb{1}_{I_{t+1/2}}) &= \gamma_t^2 \mathbb{E} \left(\langle U_{t+1/2}, x_{t+1/2} - x^* \rangle^2 \mathbb{1}_{I_{t+1/2}} \right) \\ &\leq \gamma_t^2 \mathbb{E} \left(\|U_{t+1/2}\|_*^2 \|x_{t+1/2} - x^*\|^2 \mathbb{1}_{I_{t+1/2}} \right) \\ &\leq \gamma_t^2 r^2 \mathbb{E} \left(\|U_{t+1/2}\|_*^2 \mathbb{1}_{I_{t+1/2}} \right), \end{aligned}$$

where we used that $I_{t+1/2} \subset \{x_{t+1/2} \in \mathcal{U}\}$. Next, by the law of total expectation and since $I_{t+1/2}$ is $\mathcal{F}_{t+1/2}$ measurable,

$$\begin{aligned} \mathbb{E}(\xi_{t+1}^2 \mathbb{1}_{I_{t+1/2}}) &\leq \gamma_t^2 r^2 \mathbb{E} \left(\mathbb{E} \left[\|U_{t+1/2}\|_*^2 \mid \mathcal{F}_{t+1/2} \right] \mathbb{1}_{I_{t+1/2}} \right) \\ &\leq \gamma_t^2 r^2 \sigma^2 \mathbb{P}(I_{t+1/2}) \\ &\leq \gamma_t^2 r^2 \sigma^2 \mathbb{P}(I_1). \end{aligned}$$

Combining these bounds, we get,

$$\begin{aligned} &\sum_{t \in \frac{1}{2}\mathbb{N}, t \geq 1} \mathbb{E} \left(\left(\xi_{t+1/2}^2 + \chi_{t+1/2} \right) \mathbb{1}_{I_t} \right) \\ &= \sum_{t \in \mathbb{N}, t \geq 1} \mathbb{E} \left(\xi_{t+1}^2 \mathbb{1}_{I_{t+1/2}} \right) + \sum_{t \in \mathbb{N}, t \geq 1} \mathbb{E} \left(\chi_{t+1/2}^2 \mathbb{1}_{I_t} \right) + \sum_{t \in \mathbb{N}, t \geq 1} \mathbb{E} \left(\chi_{t+1}^2 \mathbb{1}_{I_{t+1/2}} \right) \\ &\leq \sigma^2 (8 + r^2) \mathbb{P}(I_1) \sum_{t=1}^{+\infty} \gamma_t^2 \\ &\leq \underbrace{\frac{\sigma^2 (8 + r^2) \Gamma}{\epsilon}}_{=\delta} \epsilon \mathbb{P}(I_1), \end{aligned}$$

which corresponds to the statement.

Hence, [Hsieh et al. \(2020, Lem. F.1\)](#) gives that,

$$\mathbb{P} \left[\bigcap_{t \in \frac{1}{2}\mathbb{N}, t \geq 1} I_t \mid I_1 \right] \geq 1 - \delta,$$

which implies our statement since $\bigcap_{t \in \frac{1}{2}\mathbb{N}, t \geq 1} I_t \subset \{\forall t \geq \frac{1}{2}, x_t \in \mathcal{U}\}$. ■

F.5 CONVERGENCE OF OPTIMISTIC MIRROR DESCENT

Proof. We prove [Proposition 7.14](#).

1. Start from [Lemma F.3.2](#) which gives us the descent inequality

$$\begin{aligned} D(x^*, x_{t+1}) + \phi_{t+1} &\leq D(x^*, x_t) + (1 - \gamma_t \mu) \phi_t - \gamma_t \langle g_{t+1/2}, x_{t+1/2} - x^* \rangle \\ &\quad + \left(4\gamma_t^2 L^2 - \frac{1}{2} \right) \|x_{t+1/2} - x_t\|^2 + 4\gamma_t^2 (\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2) \end{aligned} \tag{F.5.1}$$

where $\phi_t = \frac{\gamma_{t-1}^2}{2} \|g_{t-1/2} - g_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = \frac{1}{2} \|x_1 - x_{1/2}\|^2 = 0$ by assumption.

2. The first part of the result comes the stochastic stability lemma [Lemma F.4.2](#) (which relies on [Hsieh et al. \(2020\)](#)), where we use that the stepsize choice is square-summable with

$$\sum_{t=1}^{+\infty} \gamma_t^2 = \gamma \leq \frac{\gamma^2}{(2\eta - 1)t_0^{2\eta-1}}. \quad (\text{F.5.2})$$

3. We now focus on proving the rates of convergence. Take $r > 0$ small enough so that $\mathcal{U} := \mathcal{B}(x^*, r) \cap \mathcal{X}$ is included in both $\mathcal{U}$ and $\mathcal{B}$.

Define the event $E_t = \{\forall 1 \leq s \leq t, x_{s+1/2} \in \mathcal{U}\}$ for $t \geq 0$. Note that, except for $t = 0$, E_t is $\mathcal{F}_{t+1/2}$ is measurable. On this event, we can apply [Lemma F.1.6](#) with $x \leftarrow x_t$ to get,

$$\langle g(x_{t+1/2}), x_{t+1/2} - x^* \rangle \geq \frac{\mu}{2} \|x_t - x^*\|^2 - \mu \|x_{t+1/2} - x_t\|^2.$$

and using the Legendre exponent of h , we get

$$\langle g(x_{t+1/2}), x_{t+1/2} - x^* \rangle \geq \frac{\mu}{K} D(x^*, x_t)^{1+\alpha} - \mu \|x_{t+1/2} - x_t\|^2.$$

4. Combining this with the descent inequality above ([F.5.1](#)), and the fact that $K \geq 1$,

$$\begin{aligned} & (D(x^*, x_{t+1}) + \phi_{t+1}) \mathbb{1}_{E_t} \\ & \leq (D(x^*, x_t) - \frac{\mu\gamma_t}{K} D(x^*, x_t)^{1+\alpha} + (1 - \frac{\gamma_t\mu}{K})\phi_t) \mathbb{1}_{E_t} - \gamma_t \langle U_{t+1/2}, x_{t+1/2} - x^* \rangle \mathbb{1}_{E_t} \\ & \quad + \left(4\gamma_t^2 L^2 + \mu\gamma_t - \frac{1}{2}\right) \|x_{t+1/2} - x_t\|^2 \mathbb{1}_{E_t} + 4\gamma_t^2 (\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2) \mathbb{1}_{E_t}. \end{aligned}$$

Now, the sequence of events $(E_s)_{s \geq 0}$ is non-increasing, so $\mathbb{1}_{E_t} \leq \mathbb{1}_{E_{t-1}} \leq 1$. As a consequence,

$$\begin{aligned} & \left(D(x^*, x_t) - \frac{\mu\gamma_t}{K} D(x^*, x_t)^{1+\alpha} + (1 - \frac{\gamma_t\mu}{K})\phi_t \right) \mathbb{1}_{E_t} \\ & \leq \left(D(x^*, x_t) - \frac{\mu\gamma_t}{K} D(x^*, x_t)^{1+\alpha} + (1 - \gamma_t\mu)\phi_t \right) \mathbb{1}_{E_{t-1}}. \end{aligned}$$

Note that the term between parenthesis is always non-negative even when $\alpha > 0$.

Moreover, by the choice of γ , $4\gamma_t^2 L^2 + \mu\gamma_t - \frac{1}{2} \leq 0$. Therefore, the descent inequality can be simplified to give,

$$\begin{aligned} & (D(x^*, x_{t+1}) + \phi_{t+1}) \mathbb{1}_{E_t} \\ & \leq \left(D(x^*, x_t) - \frac{\mu\gamma_t}{K} D(x^*, x_t)^{1+\alpha} + (1 - \frac{\gamma_t\mu}{K})\phi_t \right) \mathbb{1}_{E_{t-1}} - \gamma_t \langle U_{t+1/2}, x_{t+1/2} - x^* \rangle \mathbb{1}_{E_t} \\ & \quad + 4\gamma_t^2 (\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2). \end{aligned}$$

We now take the expectation of this inequality. For this, note that, as E_t is $\mathcal{F}_{t+1/2}$ measurable,

$$\mathbb{E} [\langle U_{t+1/2}, x_{t+1/2} - x^* \rangle \mathbb{1}_{E_t}] = \mathbb{E} [\mathbb{E} [\langle U_{t+1/2}, x_{t+1/2} - x^* \rangle \mathbb{1}_{E_t} \mid \mathcal{F}_{t+1/2}]]$$

$$\begin{aligned}
&= \mathbb{E} \left[\langle \mathbb{E} [U_{t+1/2} \mid \mathcal{F}_{t+1/2}], x_{t+1/2} - x^* \rangle \mathbb{1}_{E_t} \right] \\
&= 0.
\end{aligned}$$

As a consequence, and using the assumption on the noise,

$$\begin{aligned}
&\mathbb{E} \left[(D(x^*, x_{t+1}) + \phi_{t+1}) \mathbb{1}_{E_t} \right] \\
&\leq \mathbb{E} \left[\left(D(x^*, x_t) - \frac{\mu \gamma_t}{K} D(x^*, x_t)^{1+\alpha} + \left(1 - \frac{\gamma_t \mu}{K} \right) \phi_t \right) \mathbb{1}_{E_{t-1}} \right] + 8\gamma_t^2 \sigma^2.
\end{aligned} \tag{F.5.3}$$

5. The final step will be to study the sequence (a_t) defined by $a_t = \mathbb{E} \left[(D(x^*, x_t) + \phi_t) \mathbb{1}_{E_{t-1}} \right]$ for all $t \geq 1$. Indeed, a bound on $a_t = \mathbb{E} \left((D(x^*, x_t) + \phi_t) \mathbb{1}_{E_{t-1}} \right)$ implies a bound on the quantity of interest,

$$\mathbb{E} (D(x^*, x_t) + \phi_t \mid E_{\mathcal{U}}) .$$

As $E_{\mathcal{U}} \subset E_t$,

$$\begin{aligned}
\mathbb{E} [D(x^*, x_t) + \phi_t \mid E_{\mathcal{U}}] &= \frac{\mathbb{E} [(D(x^*, x_t) + \phi_t) \mathbb{1}_E]}{\mathbb{P}[E_{\mathcal{U}}]} \\
&\leq \frac{\mathbb{E} [(D(x^*, x_t) + \phi_t) \mathbb{1}_{E_t}]}{\mathbb{P}[E_{\mathcal{U}}]}
\end{aligned}$$

Now, the inequality (F.5.3) above can be rewritten as follows,

$$a_{t+1} \leq a_t - \frac{\mu \gamma_t}{K} \mathbb{E} \left[\left(D(x^*, x_t)^{1+\alpha} + \phi_t \right) \mathbb{1}_{E_{t-1}} \right] + 8\gamma_t^2 \sigma^2. \tag{F.5.4}$$

6. Now, the behavior of this sequence depends heavily on the Legendre exponent α :

- If $\alpha = 0$, (F.5.4) becomes

$$a_{t+1} \leq \left(1 - \frac{\mu \gamma}{K(t + t_0)^\eta} \right) a_t + \frac{8\gamma^2 \sigma^2}{(t + t_0)^{2\eta}} .$$

Then, for $\eta = 1$, Lemma F.2.1 can be applied with the additional assumption that $\gamma > \frac{K}{\mu}$ to get that for any $T \geq 1$, $a_T = \mathcal{O}(1/T)$.

For $\frac{1}{2} < \eta < 1$, Lemma F.2.2 can be directly apply to obtain $a_T = \mathcal{O}(1/T^\eta)$.

- If $\alpha > 0$, a little more work has to be done before on (F.5.4) before concluding. First, (F.5.4) implies that for any $t \geq 1$, $a_t \leq a_0 + 8\sigma^2 \sum_{i=1}^t \gamma_i^2$ and since we assume $D(x^*, x_1) + \phi_1 = D(x^*, x_1) \leq \frac{r^2}{12}$ we have that

$$\begin{aligned}
\mathbb{E} [\phi_t \mathbb{1}_{E_{t-1}}] &\leq \mathbb{E} [(D(x^*, x_t) + \phi_t) \mathbb{1}_{E_{t-1}}] = a_t \\
&\leq \mathbb{E} [(D(x^*, x_1) + \phi_1) \mathbb{1}_{E_0}] + 8\sigma^2 \sum_{t=1}^{\infty} \gamma_t^2 \\
&\leq \underbrace{\frac{r^2}{12} + \frac{8\sigma^2 \gamma^2}{(2\eta - 1)t_0^{2\eta-1}}}_{:=\Phi} .
\end{aligned}$$

As a consequence, $-\mathbb{E}[\phi_t \mathbb{1}_{E_{t-1}}] \leq -\mathbb{E}[\phi_t \mathbb{1}_{E_{t-1}}]^{1+\alpha} / \Phi^\alpha$, and thus (F.5.4) gives us

$$\begin{aligned} a_{t+1} &\leq a_t - \frac{\mu\gamma_t}{K} \mathbb{E} \left[\left(D(x^*, x_t)^{1+\alpha} + \frac{1}{\Phi^\alpha} \phi_t^{1+\alpha} \right) \mathbb{1}_{E_{t-1}} \right] + 8\gamma_t^2 \sigma^2 \\ &\leq a_t - \frac{\mu\gamma_t}{2^\alpha \max(1, \Phi^\alpha) K} \mathbb{E} \left[(D(x^*, x_t) + \phi_t)^{1+\alpha} \mathbb{1}_{E_{t-1}} \right] + 8\gamma_t^2 \sigma^2 \\ &\leq a_t - \frac{\mu\gamma_t}{2^\alpha \max(1, \Phi^\alpha) K} a_t^{1+\alpha} + 8\gamma_t^2 \sigma^2 \end{aligned}$$

where the second inequality comes from the fact that $(x + y)^{1+\alpha} / 2^\alpha \leq x^{1+\alpha} + y^{1+\alpha}$ for positive x, y and the last one from Jensen inequality applied to the convex function $x \mapsto x^{1+\alpha}$.

We will now apply one of Lemmas F.2.4 and F.2.5 to the sequence (a_t) . To make this step clear, let us introduce the same notations as these lemmas. Define

$$q := \frac{\mu\gamma}{2^\alpha \max(1, \Phi^\alpha) K} \quad \text{and} \quad q' := 8\sigma^2 \gamma^2.$$

With these notations, the descent inequality can be rewritten as,

$$a_{t+1} \leq a_t - \frac{q}{(t + t_0)^\eta} a_t^{1+\alpha} + \frac{q'}{(t + t_0)^{2\eta}}.$$

Both Lemmas F.2.4 and F.2.5 require that,

$$q' q^{1/\alpha} \leq c(\eta, \alpha) \iff \gamma^{2+\frac{1}{\alpha}} \leq \frac{c(1, \alpha)}{4\sigma^2(\mu/K)^{1/\alpha}} \max(1, \Phi).$$

Finally, we distinguish two cases.

- If $\eta \geq \frac{1+\alpha}{1+2\alpha}$, we apply Lemma F.2.4 to get that, for any $T \geq 1$,

$$a_T \leq \frac{a_1 + b}{\left(1 + \alpha(a_1 + b)^\alpha 2^{-\alpha} \sum_{t=1}^{T-1} \frac{q}{(t+t_0)^\eta}\right)^{1/\alpha}},$$

where $b = \left(\frac{1-2^{1-2\eta}}{(1+\alpha)q}\right)^{\frac{1}{\alpha}}$. Using asymptotic notations, this simply means that,

$$a_T = \mathcal{O}^* \left(\sum_{t=1}^{T-1} \frac{1}{(t+t_0)^\eta} \right)^{-1/\alpha}.$$

The final bound on a_T now comes from the fact that,

$$\sum_{t=1}^{T-1} \frac{1}{(t+t_0)^\eta} = \begin{cases} \Theta(T^{1-\eta}) & \text{if } \eta < 1 \\ \Theta(\log T) & \text{if } \eta = 1. \end{cases}$$

- If $\eta \leq \frac{1+\alpha}{1+2\alpha}$, we apply Lemma F.2.5 to get that, for any $T \geq 1$,

$$a_T \leq \frac{a_1}{\left(1 + \alpha a_1^\alpha \sum_{t=1}^{T-1} \frac{q}{(t+t_0)^\eta}\right)^{1/\alpha}} + \frac{1}{((1+\alpha)q)^{\frac{1}{\alpha}} (t+t_0)^\eta}.$$

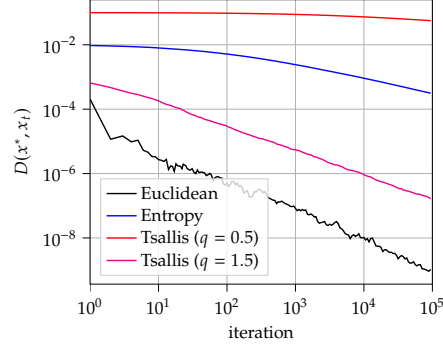


Figure 27: Convergence of OMD with different DGFs in terms of Bregman divergence to the solution.

In other words, the sequence (a_t) satisfies, as $\eta < 1$,

$$a_T = \mathcal{O}^* \frac{1}{T^{(1-\eta)/\alpha}} + \mathcal{O}^* \frac{1}{T^\eta}.$$

■

F.6 NUMERICAL EXPERIMENTS

We describe here the setup used in the illustration of [Theorem 7.13](#) and provide additional plots.

We consider a simple 1D example with $\mathcal{X} = [0, +\infty)$ and $g(x) = x$; the solution of the associated variational inequality is thus $x^* = 0$. It is direct to see that g is $L = 1$ Lipschitz continuous and $\mu = 1$ strongly monotone. We consider the three Bregman regularizers of our running examples on $\mathcal{X} = [0, +\infty)$:

- *Euclidean projection* ([Example 7.1](#)): $h(x) = x^2/2$ for which $D(x^*, x) = x^2/2$, and $\beta = 0$;
- *Negative entropy* ([Example 7.2](#)): $h(x) = x \log x$ for which $D(x^*, x) = x$ and $\beta = 1/2$ at x^* ;
- *Tsallis entropy* ([Example 7.3](#)): $h(x) = -\frac{1}{q(1-q)}x^q$ for which $D(x^*, x) = x^q/q$ and $\beta = \max(0, 1 - q/2)$ at x^* . To show different behaviors we consider $q = 0.5$ ($\beta = 0.75$) and $q = 1.5$ ($\beta = 0.25$).

For each of these regularizers, we run OMD initialized with $x_1 = x_{1/2} = 0.1$, (U_t) an i.i.d. Gaussian noise process with mean 0 and variance $\sigma^2 = 10^{-4}$. We choose our stepsize sequence as $\gamma_t = \frac{1}{t^\eta}$ with η as prescribed by [Theorem 7.13](#). We display the results averaged over 100 runs.

We find out that the observed rates match the theory:

Regularizer	β	γ_t	Theoretical rate	Observed rate (by regression)
Euclidean	0	$\frac{1}{t}$	$\frac{1}{t}$	$\frac{1}{t^{0.99}}$
Entropy	0.5	$\frac{1}{t^{0.5+\epsilon}}$	$\frac{1}{t^{0.5+\epsilon}}$	$\frac{1}{t^{0.48}}$
Tsallis ($q = 0.5$)	0.75	$\frac{1}{t^{0.5+\epsilon}}$	$\frac{1}{t^{0.1666}}$	$\frac{1}{t^{0.13}}$
Tsallis ($q = 1.5$)	0.25	$\frac{1}{t^{0.75}}$	$\frac{1}{t^{0.75}}$	$\frac{1}{t^{0.71}}$

LIST OF PUBLICATIONS

This appendix provides a list of all the publications from the author, see § 1.6 for details and comments.

- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. What is the long-run distribution of stochastic gradient descent? A large deviations analysis. In *ICML 2024*, 2024b
- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The global convergence time of stochastic gradient descent in non-convex landscapes: Sharp estimates via large deviations. In *ICML*, 2025a
- Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Exact Generalization Guarantees for (Regularized) Wasserstein Distributionally Robust Models. In *NeurIPS*, 2023c
- Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for Wasserstein distributionally robust optimization. *ESAIM: COCV*, 2023a
- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The rate of convergence of Bregman proximal methods: Local geometry vs. regularity vs. sharpness. *SIAM Journal on Optimization*, September 2024a
- Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *COLT 2021*, 2021a
- Florian Vincent, Waïss Azizian, Franck Iutzeler, and Jérôme Malick. skwdro: a library for Wasserstein distributionally robust machine learning. *Journal of Machine Learning Research*, 2026
- Waïss Azizian and Ali Hasan. How does the pretraining distribution shape in-context learning? Task selection, generalization, and robustness. *arXiv preprint arXiv: 2510.01163*, 2025
- Simon Weissmann, Sara Klein, Waïss Azizian, and Leif Döring. Almost sure convergence of stochastic gradient methods under gradient domination. *Transactions on Machine Learning Research*, 2025
- Waïss Azizian, Michael Kirchhof, Eugene Ndiaye, Louis Béthune, Michal Klein, Pierre Ablin, and marco cuturi. The geometries of truth are orthogonal across tasks. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025b
- Waïss Azizian and Marc Lelarge. Expressive power of invariant and equivariant graph neural networks. In *ICLR*, 2021
- Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *AISTATS*, 2020a

- Adam Ibrahim, Waiss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International conference on machine learning*, pages 4583–4593. PMLR, 2020
- Waiss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *AISTATS*, 2020b

COLOPHON

This manuscript was typeset with $\text{\LaTeX} 2_{\epsilon}$ using Hermann Zapf’s Palatino type face (the actual Type 1 PostScript fonts used were URW Palladio L and FPL). The monospaced text (hyperlinks, etc.) was typeset in *Bera Mono*, originally developed by Bitstream, Inc. as “Bitstream Vera” (with Type 1 PostScript fonts by Malte Rosenau and Ulrich Dirr).

The typographic style of this dissertation was inspired by the authoritative genius of Bringhurst’s *Elements of Typographic Style*, ported to $\text{\LaTeX}$ by André Miede, the original designer of the [classicthesis](#) template. Any unsightly deviations from these works should be attributed solely to the author’s (not always successful) efforts to conform to the awkward A4 paper size.

Stochastic optimization algorithms in machine learning: dynamics, convergence, generalization

© Waïss Azizian 2026

Grenoble, June 16, 2026

Waïss Azizian