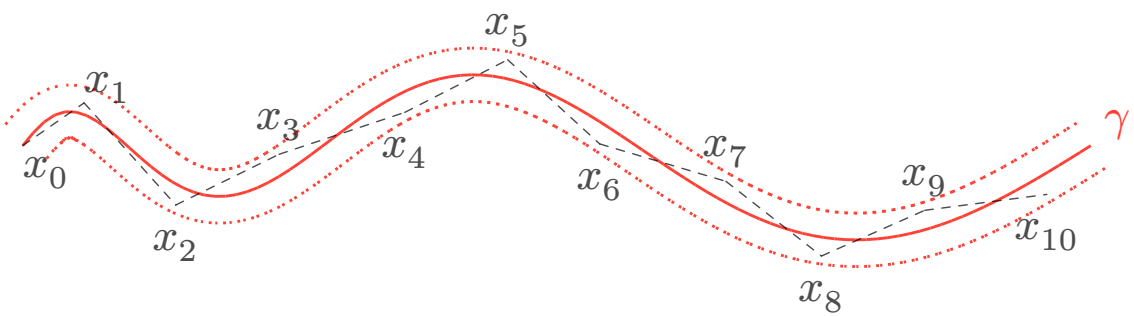
Stochastic Optimization Algorithms in Machine Learning

Dynamics, Convergence, Generalization



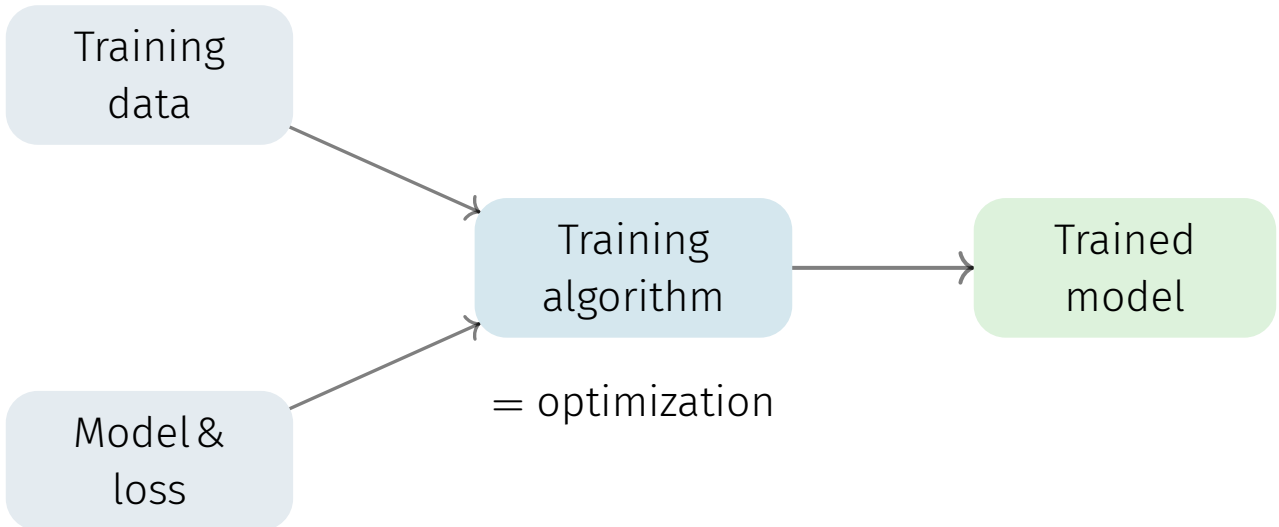
PhD defense

June 16, 2026

W. Azizian

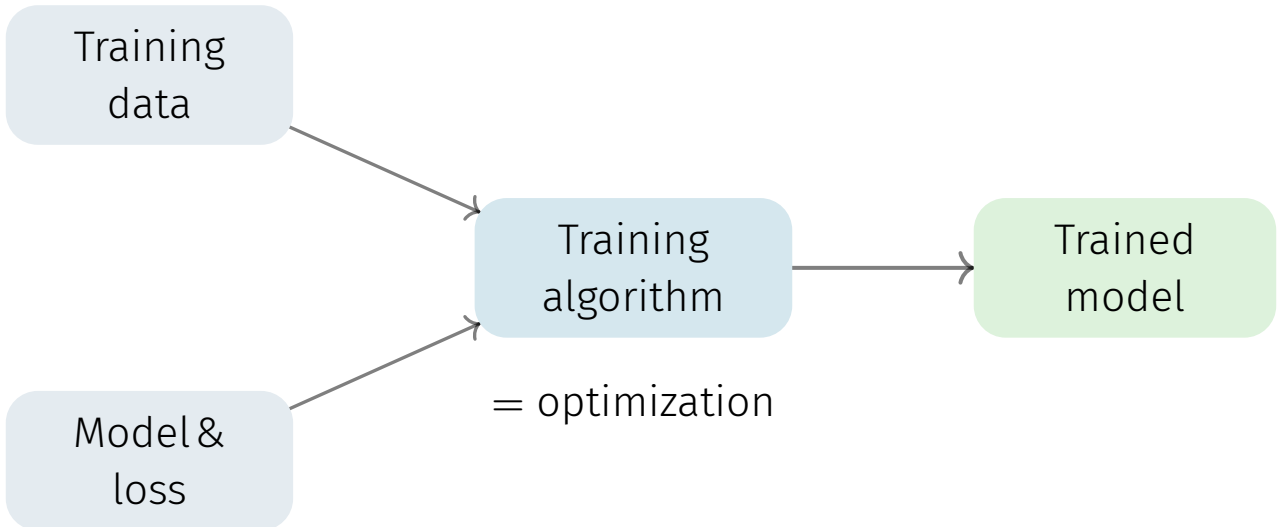
ML through the lens of optimization

"given data, find the model that best represents the data"



ML through the lens of optimization

"given data, find the model that best represents the data"



Thread 1
Non-convexity
Where does training converge?

Thread 2
Generalization
Will the model perform well on new data?

Thread 3
Constraints
How do we optimize over intricate geometries?

This Thesis

Thread 1 — ICML ‘24, ICML ‘25

Stochastic optimization with non-convex losses

What is the long-run behavior of these algorithms?

Thread 2 — COCV ‘23, NeurIPS ‘23

Robust optimization with Wasserstein uncertainty sets

How can we control the test error by carefully designing the training objective?

Thread 3 — COLT ‘22, SIOPT ‘24

Last-iterate convergence of Bregman proximal methods

How can we explain the different behaviors across divergences?

This Thesis

Thread 1 — ICML ‘24, ICML ‘25

Stochastic optimization with non-convex losses

What is the long-run behavior of these algorithms?

Thread 2 — COCV ‘23, NeurIPS ‘23

Robust optimization with Wasserstein uncertainty sets

How can we control the test error by carefully designing the training objective?

Thread 3 — COLT ‘22, SIOPT ‘24

Last-iterate convergence of Bregman proximal methods

How can we explain the different behaviors across divergences?

On the long-run behavior of SGD in non-convex landscapes

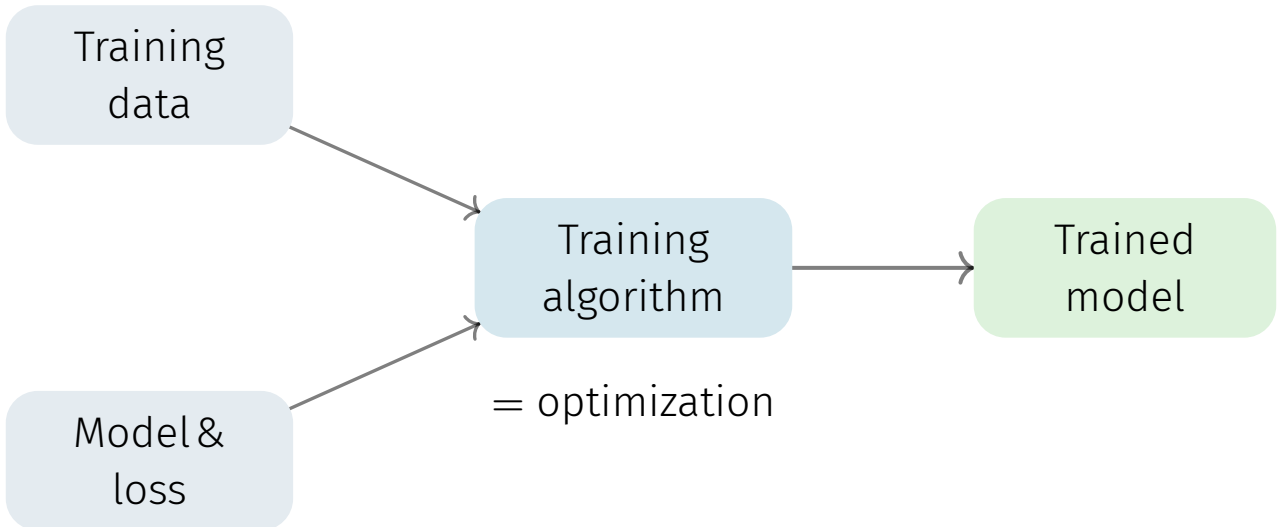
1. **Introduction** Context and informal overview of Result 1
2. **Our Approach** Essential ideas, key objects and proof sketch of Result 1
3. **Result 1** Long-run distribution of SGD
4. **Result 2** Time to reach the global minimum
5. **Perspectives** Beyond SGD

On the long-run behavior of SGD in non-convex landscapes

1. **Introduction** Context and informal overview of Result 1
2. **Our Approach** Essential ideas, key objects and proof sketch of Result 1
3. **Result 1** Long-run distribution of SGD
4. **Result 2** Time to reach the global minimum
5. **Perspectives** Beyond SGD

ML through the lens of optimization

"given data, find the model that best represents the data"



Thread 1
Non-convexity
Where does training converge?

Why non-convex landscapes?

Training of deep neural networks = stochastic optimization on a nonconvex loss function

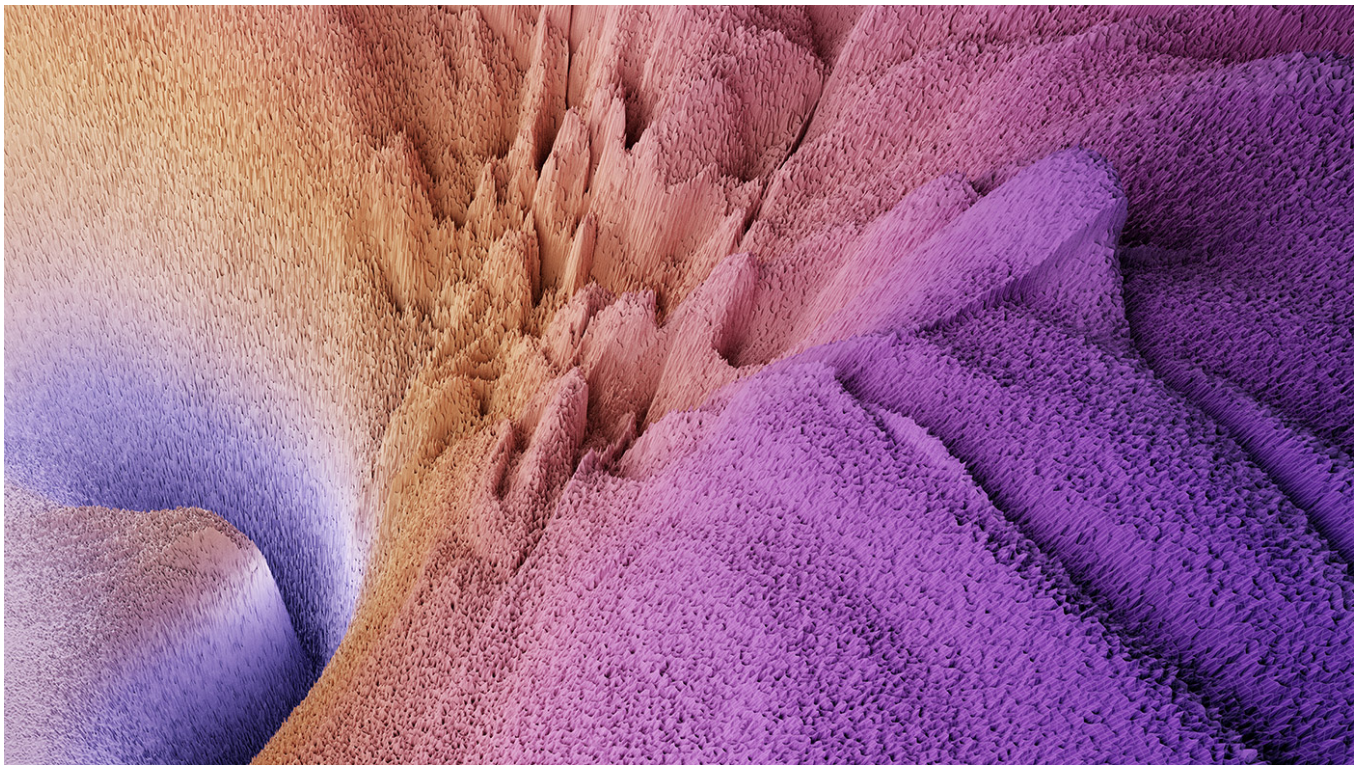


Image credit: losslandscape.com

Why non-convex landscapes?

Training of deep neural networks = stochastic optimization on a nonconvex loss function

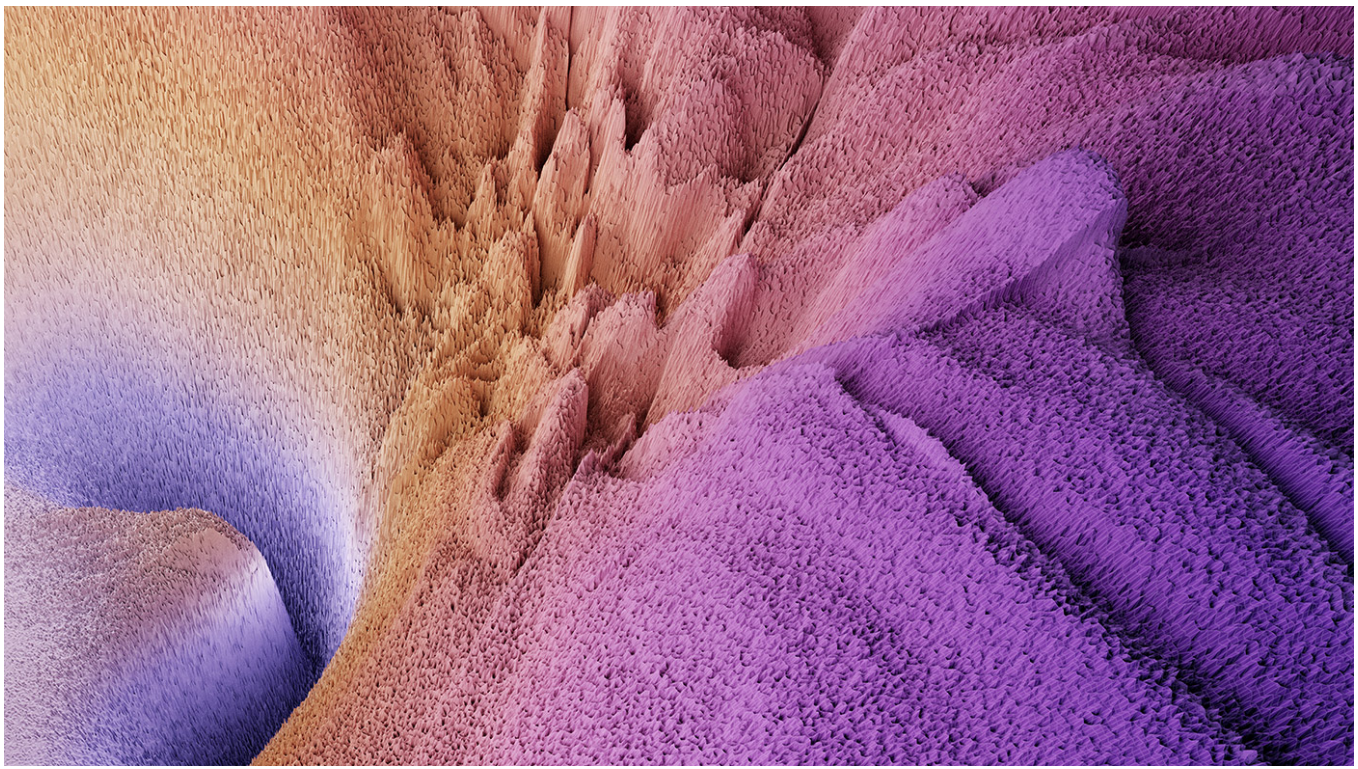


Image credit: losslandscape.com

Q: What is the behavior of stochastic optimization methods on a non-convex loss?

Stochastic Gradient Descent

For $f : \mathbb{R}^d \to \mathbb{R}$ non-convex, eg. loss of model with parameters x,

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x)$$

Stochastic Gradient Descent (SGD): with *constant* step size $\eta > 0$

$$x_{t+1} = x_t - \eta \left[\nabla f(x_t) + Z(x_t; \omega_t) \right]$$

step size — zero-mean noise

Stochastic Gradient Descent

For $f : \mathbb{R}^d \to \mathbb{R}$ non-convex, eg. loss of model with parameters x,

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x)$$

Stochastic Gradient Descent (SGD): with *constant* step size $\eta > 0$

$$x_{t+1} = x_t - \underbrace{\eta}_{\text{step size}} \Big[\nabla f(x_t) + \underbrace{Z(x_t; \omega_t)}_{\text{zero-mean noise}} \Big]$$

Finite-sum problems / Empirical risk minimization:

Consider $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ and at each iteration, sample i_t,

$$Z(x_t; \omega_t) = \nabla f(x_t) - \nabla f_{i_t}(x_t)$$

Stochastic Gradient Descent

For $f : \mathbb{R}^d \to \mathbb{R}$ non-convex, eg. loss of model with parameters x,

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x)$$

Stochastic Gradient Descent (SGD): with *constant* step size $\eta > 0$

$$x_{t+1} = x_t - \underbrace{\eta}_{\text{step size}} \Big[\nabla f(x_t) + \underbrace{Z(x_t; \omega_t)}_{\text{zero-mean noise}} \Big]$$

Q: What is the long-run behavior of SGD?

Stochastic Gradient Descent

For $f : \mathbb{R}^d \to \mathbb{R}$ non-convex, eg. loss of model with parameters x,

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x)$$

Stochastic Gradient Descent (SGD): with *constant* step size $\eta > 0$

$$x_{t+1} = x_t - \underbrace{\eta}_{\text{step size}} \Big[\nabla f(x_t) + \underbrace{Z(x_t; \omega_t)}_{\text{zero-mean noise}} \Big]$$

Q: What is the long-run behavior of SGD?

→ **Q1:** Where are the iterates most likely to go?

→ **Q2:** How much time to get there?

What is known on SGD?

Stochastic Gradient Descent (SGD): with *constant* step size $\eta > 0$

$$x_{t+1} = x_t - \eta \left[\nabla f(x_t) + Z(x_t; \omega_t) \right]$$

SGD with constant step size:

- f strongly convex: SGD converges near the minimizer (Polyak, 1987)
- f convex: average of SGD iterates (almost) optimal (Nemirovskii and Yudin (1978, 1983))
- f nonconvex:
  - In average, close to criticality (Lan, 2012)

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \right] = \mathcal{O}\left(\frac{1}{\sqrt{T}} \right)$$

  - With probability 1, SGD is not stuck in (strict) saddle points (Pemantle, 1990; Mertikopoulos et al., 2020)

→ **Q1:** Where are the iterates most likely to go?

→ **Q2:** How much time to get to the global minimum?

What is known on SGD?

Stochastic Gradient Descent (SGD): with *constant* step size $\eta > 0$

$$x_{t+1} = x_t - \eta \Big[\nabla f(x_t) + Z(x_t; \omega_t) \Big]$$

What we are not doing:

- Stochastic Approximation:

$$x_{t+1} = x_t - \eta_t [\nabla f(x_t) + Z(x_t; \omega_t)] \text{ with } \eta_t \propto \frac{1}{t^{0.5+\varepsilon}}$$

  Convergence to local minima (Bertsekas & Tsitsiklis, 2000) but does not correspond to practice.

- Sampling (MCMC, Langevin): to sample from e^{-f/σ^2}

$$x_{t+1} = x_t - \eta \nabla f(x_t) + \sqrt{2\eta} \mathcal{N}(0, \sigma^2)$$

  Convergence of the distribution of the iterates (Raginsky et al., 2017) but scaling of the noise differs from SGD

- Continuous-time limit (Gradient flow, SDE):

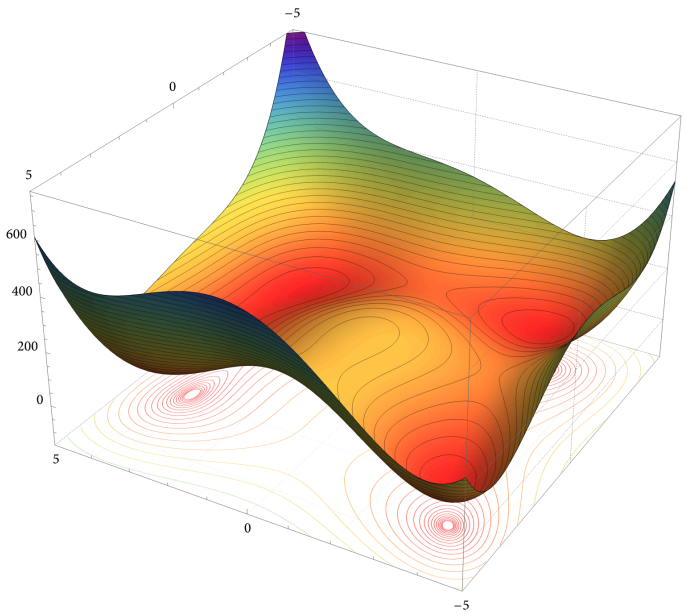
$$dX_t = -\nabla f(X_t) dt + \sqrt{\eta \operatorname{cov}(Z(X_t; \cdot))} dW_t$$

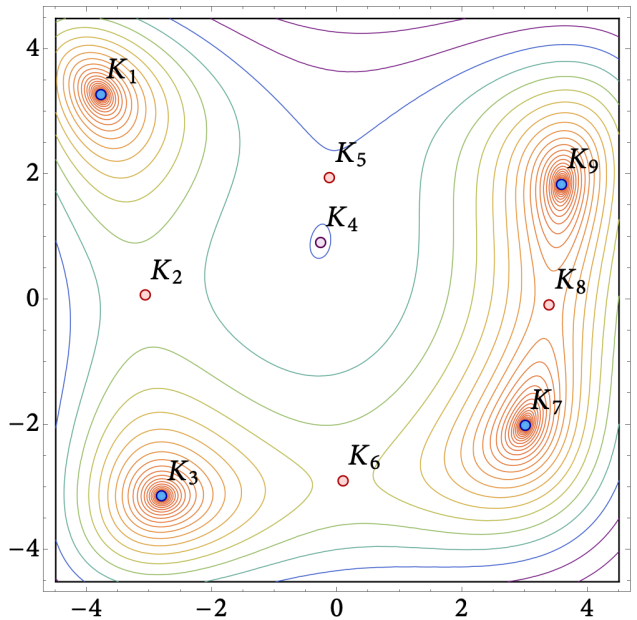
  Provable approximation of SGD (Li et al., 2017) but only on finite time horizons

$\Rightarrow$ Cannot analyze the long-run behavior of SGD

Example: Himmelblau function

Himmelblau function



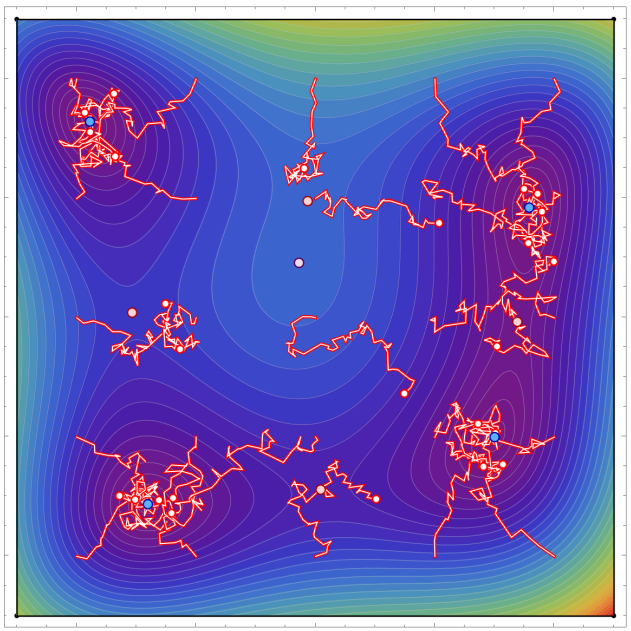


In our work, critical set of f:

$$\mathrm{crit}(f) := \{x : \nabla f(x) = 0\} = K_1 \cup K_2 \cup \ldots \cup K_p$$

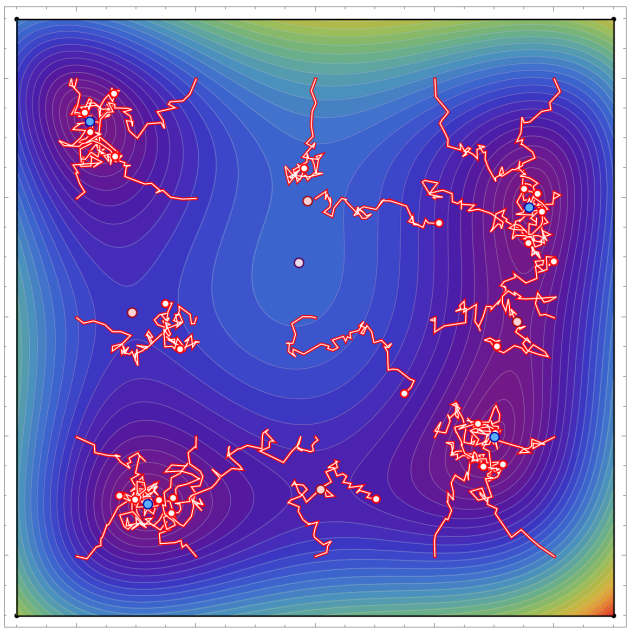
where K_i (smoothly) connected components

Example: SGD on Himmelblau function

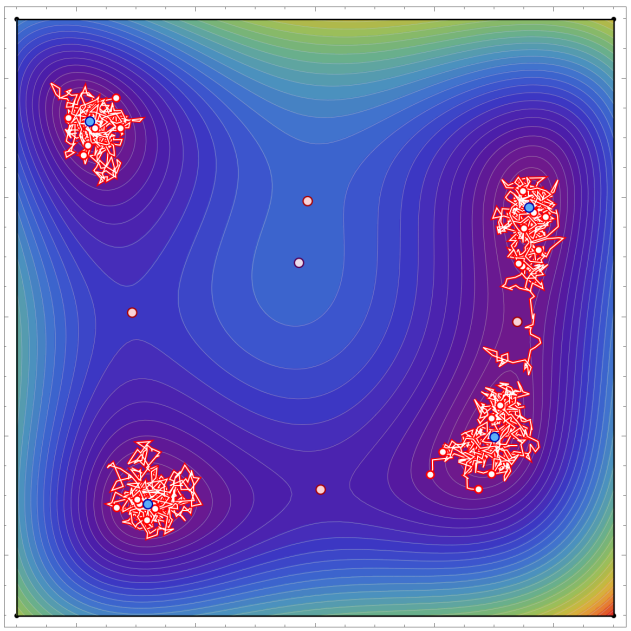


$t = 50$

Example: SGD on Himmelblau function

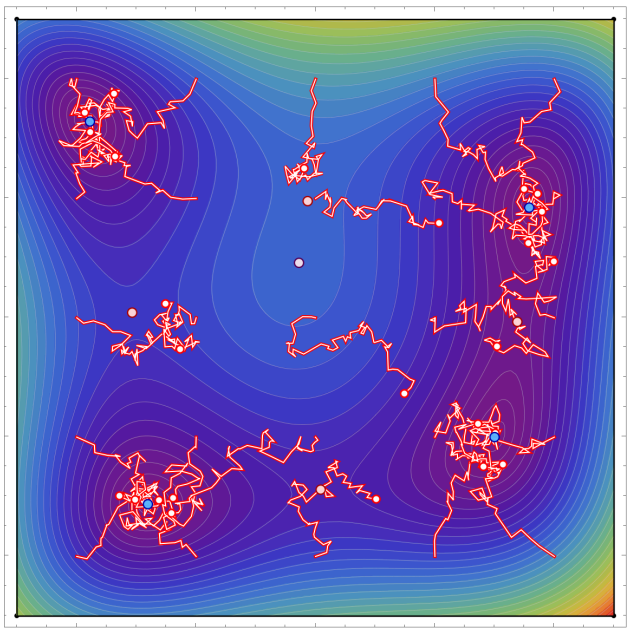


$t = 50$

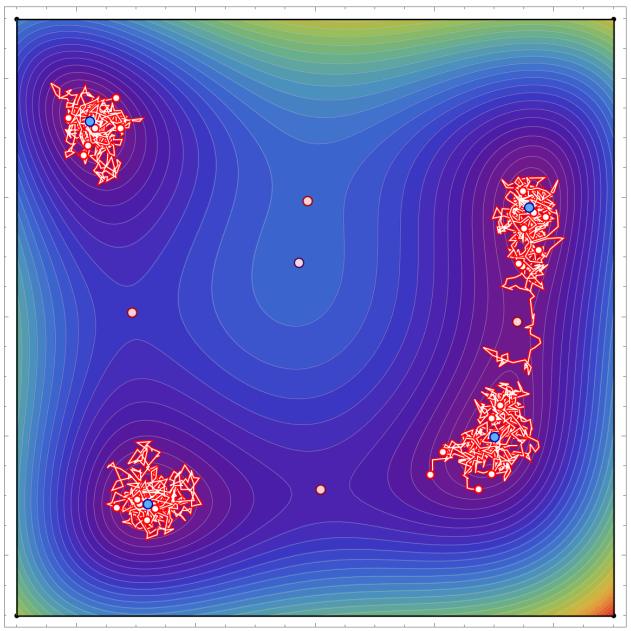


$t = 200$

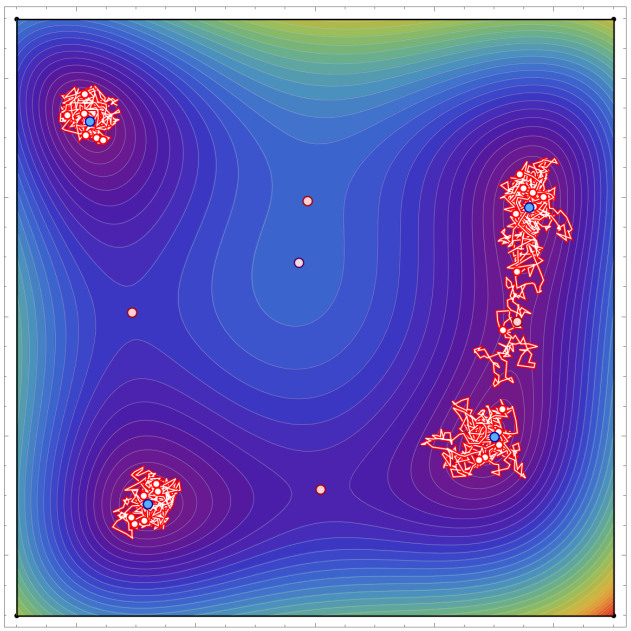
Example: SGD on Himmelblau function



$t = 50$



$t = 200$



$t = 500$

Result 1: Asymptotic distribution of SGD

SGD with *constant* step size = Markov Chain

$$x_{t+1} = x_t - \eta\Big[\nabla f(x_t) + Z(x_t;\omega_t)\Big]$$

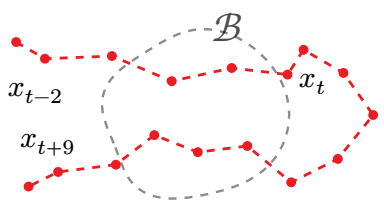
Invariant measure: *probability measure* μ_∞ *such that*

$$x_t \sim \mu_\infty \qquad \Rightarrow \qquad x_{t+1} \sim \mu_\infty$$

Invariant measures = long-run average behavior:

Invariant measures are weak-$\star$ limit points of the mean occupation measures of the iterates of SGD: for any set $\mathcal{B}$ of interest, as $n \to \infty$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} 1\{x_t \in \mathcal{B}\}\right] \approx \mu_\infty(\mathcal{B})$$



Q1: Where does the invariant measure of SGD concentrate?

Result 1: Informal statement

Recall: $\mathrm{crit}(f) := \{x : \nabla f(x) = 0\} = K_1 \cup K_2 \cup \ldots \cup K_p$ with K_i connected components

1. **Concentration near critical points:**
$$\mu_\infty(\mathrm{crit}(f)) \geq 1 - e^{-\frac{c}{\eta}}$$
2. **Saddle-point avoidance:**
$$\mu_\infty(\text{saddle point}) \ll \mu_\infty(\text{local minima})$$
3. **Boltzmann-Gibbs distribution:** for some energy levels E_i,
$$\mu_\infty(K_i) \propto \exp\left(-\frac{E_i}{\eta}\right)$$
4. **Ground state concentration:** with $K_0 := \mathrm{argmin}_i\, E_i$ minimizing energy,
$$\mu_\infty(K_0) \geq 1 - e^{-\frac{c}{\eta}}$$

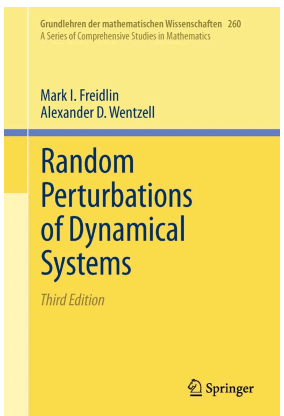
Challenges and techniques

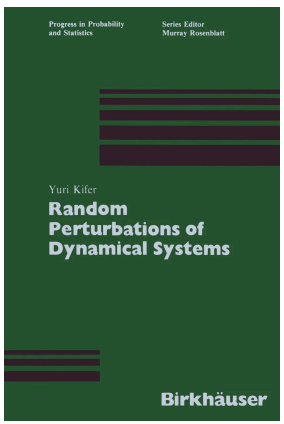
- No known approach to analyze the asymptotic distribution of SGD on non-convex problems
- We leverage large deviation theory and the theory of random perturbations of dynamical systems,
  $\rightarrow$ Estimate the probability of rare events, such as SGD escaping a local minima
- We adapt the theory of random perturbations of dynamical systems with three main challenges:
  a) Lack of compactness
  b) Realistic noise models (finite sum)
  c) Discrete-time dynamics

References

Freidlin, M. I., & Wentzell, A. D., 2012.
Random perturbations of dynamical systems.

Kifer, Y., 1988. *Random perturbations of dynamical systems*.





On the long-run behavior of SGD in non-convex landscapes

1. **Introduction** Context and informal overview of Result 1
2. **Our Approach** Essential ideas, key objects and proof sketch of Result 1
3. **Result 1** Long-run distribution of SGD
4. **Result 2** Time to reach the global minimum
5. **Perspectives** Beyond SGD

Objective and noise assumptions

Objective assumptions:

- ∇f is Lipschitz-continuous
- f is coercive: $\lim_{\|x\|\to\infty} f(x) = \lim_{\|x\|\to\infty} \|\nabla f(x)\| = +\infty$
- $\mathrm{crit}(f)$ has finitely many (smoothly) connected components:

$$\mathrm{crit}(f) = K_1 \cup K_2 \cup ... \cup K_p$$

Noise assumptions:

- Randomness ω_t is i.i.d.
- $\mathbb{E}[Z(x;\omega)] = 0$, $\mathrm{cov}(Z(x;\omega)) \succ 0$, $Z(x;\omega) = O(\|x\|)$ almost surely
- $Z(x;\omega)$ is σ sub-Gaussian:

$$\log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right] \leq \frac{\sigma^2}{2}\|v\|^2$$

Objective and noise assumptions

Objective assumptions:

- ∇f is Lipschitz-continuous
- f is coercive: $\lim_{\|x\|\to\infty} f(x) = \lim_{\|x\|\to\infty} \|\nabla f(x)\| = +\infty$
- $\mathrm{crit}(f)$ has finitely many (smoothly) connected components:

$$\mathrm{crit}(f) = K_1 \cup K_2 \cup ... \cup K_p$$

Noise assumptions:

- Randomness ω_t is i.i.d.
- $\mathbb{E}[Z(x;\omega)] = 0$, $\mathrm{cov}(Z(x;\omega)) \succ 0$, $Z(x;\omega) = O(\|x\|)$ almost surely
- $Z(x;\omega)$ is σ sub-Gaussian:

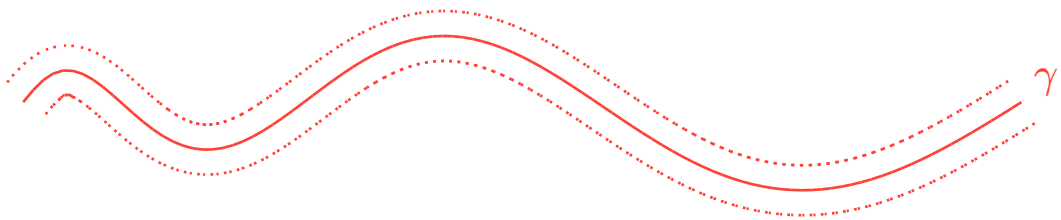
$$\log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right] \leq \frac{\sigma^2}{2}\|v\|^2$$

Regularized empirical risk minimization:

Consider $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) + \frac{\lambda}{2}\|x\|^2$ with f_i Lipschitz and smooth.

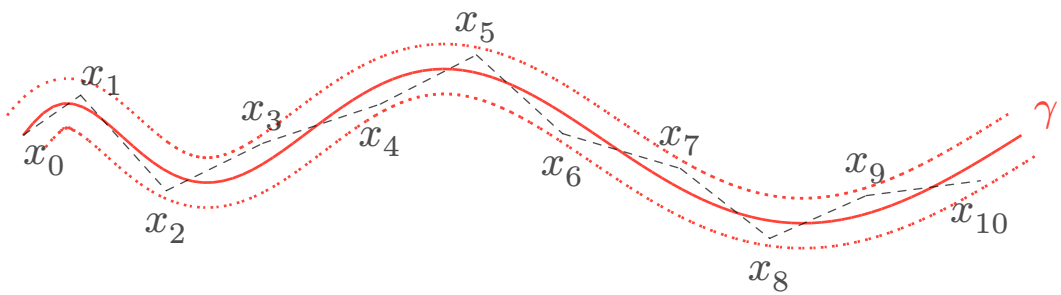
Large deviations for discrete-time SGD

Consider $\gamma : [0, T] \to \mathbb{R}^d$ continuous path in parameter space, $\mathbb{P}(\text{SGD} \approx \gamma) = ?$



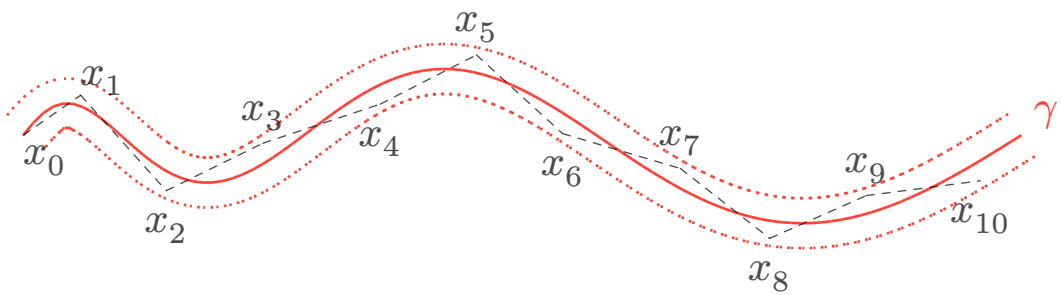
Large deviations for discrete-time SGD

Consider $\gamma : [0, T] \to \mathbb{R}^d$ continuous path in parameter space, $\mathbb{P}(\text{SGD} \approx \gamma) = ?$



Large deviations for discrete-time SGD

Consider $\gamma : [0, T] \to \mathbb{R}^d$ continuous path in parameter space, $\mathbb{P}(\text{SGD} \approx \gamma) = ?$



Proposition: *SGD admits a large deviation principle as $\eta \to 0$: for any path $\gamma : [0, T] \to \mathbb{R}^d$,*

$$\mathbb{P}(\text{SGD on } [0, T/\eta] \approx \gamma) \approx \exp\left(-\frac{\mathcal{S}_T[\gamma]}{\eta}\right) \quad \text{where } \mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt$$

Using tools from (Freidlin & Wentzell, 2012; Dupuis, 1988)

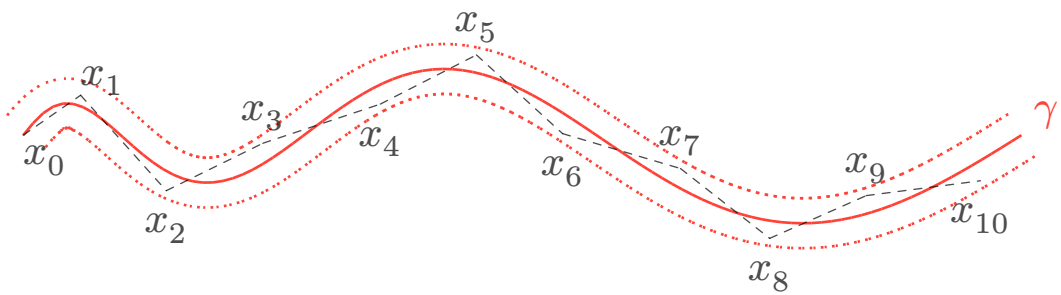
CGF. of $Z(x; \omega)$: $\mathcal{H}(x, v) = \log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right]$

Conjugate: $\mathcal{L}(x, v) = \mathcal{H}^*(x, -v - \nabla f(x))$

Action: $\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt$

Large deviations for discrete-time SGD

Consider $\gamma : [0, T] \to \mathbb{R}^d$ continuous path in parameter space, $\mathbb{P}(\text{SGD} \approx \gamma) = ?$



Proposition: *SGD admits a large deviation principle as $\eta \to 0$: for any path $\gamma : [0, T] \to \mathbb{R}^d$,*

$$\mathbb{P}(\text{SGD on } [0, T/\eta] \approx \gamma) \approx \exp\left(-\frac{\mathcal{S}_T[\gamma]}{\eta}\right) \quad \text{where } \mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt$$

Using tools from (Freidlin & Wentzell, 2012; Dupuis, 1988)

		Gaussian noise $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I_d)$
CGF. of $Z(x;\omega)$:	$\mathcal{H}(x, v) = \log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right]$	$\mathcal{H}(x, v) = \frac{\sigma^2}{2}\|v\|^2$
Conjugate:	$\mathcal{L}(x, v) = \mathcal{H}^*(x, -v - \nabla f(x))$	$\mathcal{L}(x, v) = \frac{\|v + \nabla f(x)\|^2}{2\sigma^2}$
Action:	$\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}(\gamma_t, \dot{\gamma}_t) dt$	$\mathcal{S}_T[\gamma] = \frac{1}{2\sigma^2} \int_0^T \|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2 dt$

LDP and Gradient flow

Gradient flow: path $\gamma : [0, T] \to \mathbb{R}^d$ such that

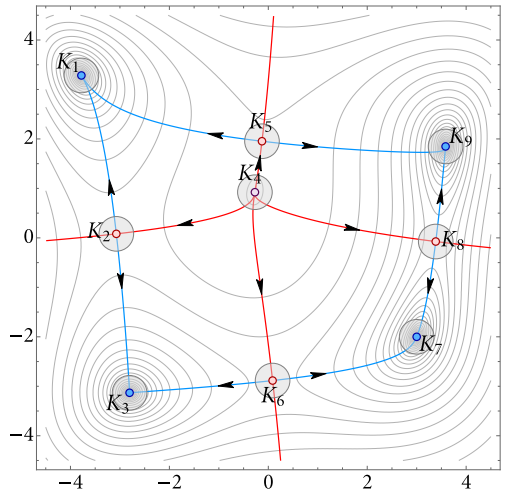
$$\dot{\gamma}_t = -\nabla f(\gamma_t)$$

Proposition:

$$\mathbb{P}(\textit{SGD on } [0, T/\eta] \approx \gamma) \approx \exp\left(-\frac{\mathcal{S}_T[\gamma]}{\eta}\right)$$

In the Gaussian case:

$$\mathcal{S}_T[\gamma] = \frac{1}{2\sigma^2} \int_0^T \|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2 dt$$



,

LDP and Gradient flow

Gradient flow: path $\gamma : [0, T] \to \mathbb{R}^d$ such that

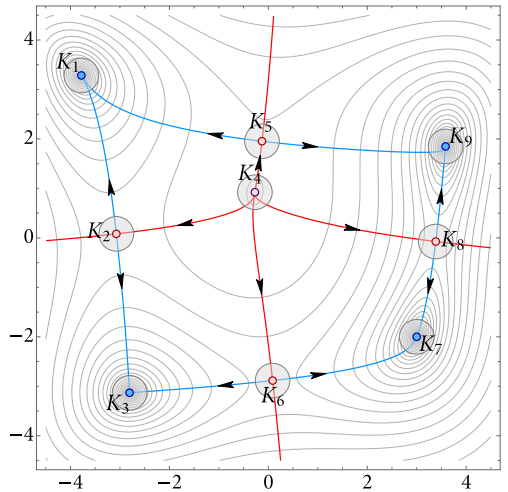
$$\dot{\gamma}_t = -\nabla f(\gamma_t)$$

Proposition:

$$\mathbb{P}(\textit{SGD on } [0, T/\eta] \approx \gamma) \approx \exp\left(-\frac{\mathcal{S}_T[\gamma]}{\eta}\right)$$

In the Gaussian case:

$$\mathcal{S}_T[\gamma] = \frac{1}{2\sigma^2} \int_0^T \|\dot{\gamma}_t + \nabla f(\gamma_t)\|^2 dt$$



,

Key observations:

- $\mathcal{S}_T[\gamma]$ = 0 iff γ is a gradient flow trajectory
- $\mathcal{S}_T[\gamma]$ quantifies how far γ is from being a gradient flow trajectory
- The farther γ is from being a gradient flow, the smaller $\mathbb{P}(\text{SGD} \approx \gamma)$

Transition between critical points

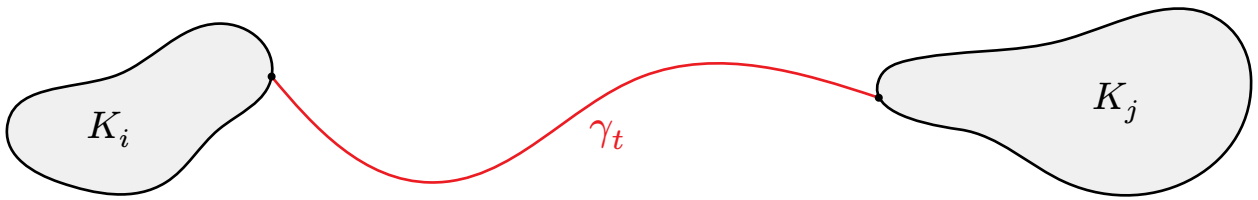
Given K_i, K_j components of critical points, what is $\mathbb{P}\big(\text{SGD transitions from } K_i \text{ to } K_j\big)$?

Transition between critical points

Given K_i, K_j components of critical points, what is $\mathbb{P}\big(\text{SGD transitions from } K_i \text{ to } K_j\big)$?

Involves the transition cost:

$$B_{i,j} = \inf\big\{\mathcal{S}_T[\gamma] \mid \gamma_0 \in K_i, \gamma_T \in K_j, T > 0\big\}$$

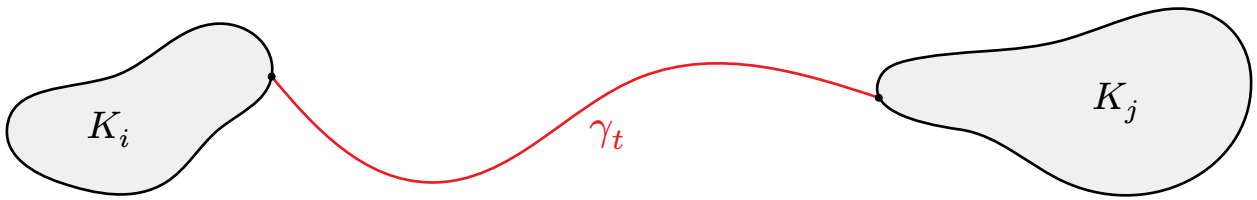


Transition between critical points

Given K_i, K_j components of critical points, what is $\mathbb{P}\big(\text{SGD transitions from } K_i \text{ to } K_j\big)$?

Involves the transition cost:

$$B_{i,j} = \inf\big\{\mathcal{S}_T[\gamma] \mid \gamma_0 \in K_i, \gamma_T \in K_j, T > 0\big\}$$



Proposition: *Transition probability from K_i to K_j: for $\eta > 0$ small enough,*

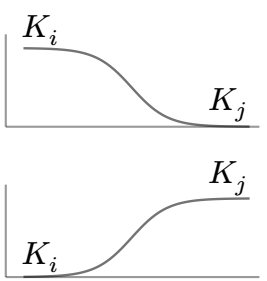
$$\mathbb{P}\big(\textit{SGD transitions from } K_i \textit{ to } K_j\big) \approx \exp\left(-\frac{B_{i,j}}{\eta}\right)$$

Assumption: $B_{i,j} < +\infty$

Key observations:

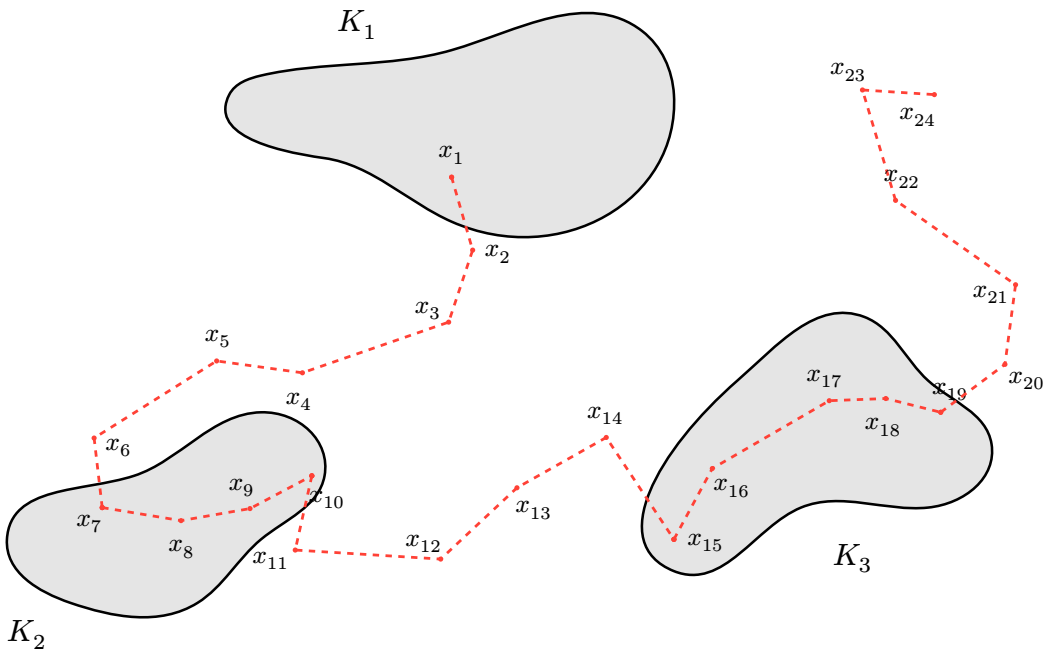
- If there is a trajectory of the gradient flow joining K_i and K_j, then $B_{i,j} = 0$
- We can show:

$$B_{i,j} \geq \frac{2\big(f(K_j) - f(K_i)\big)}{\sigma^2}$$



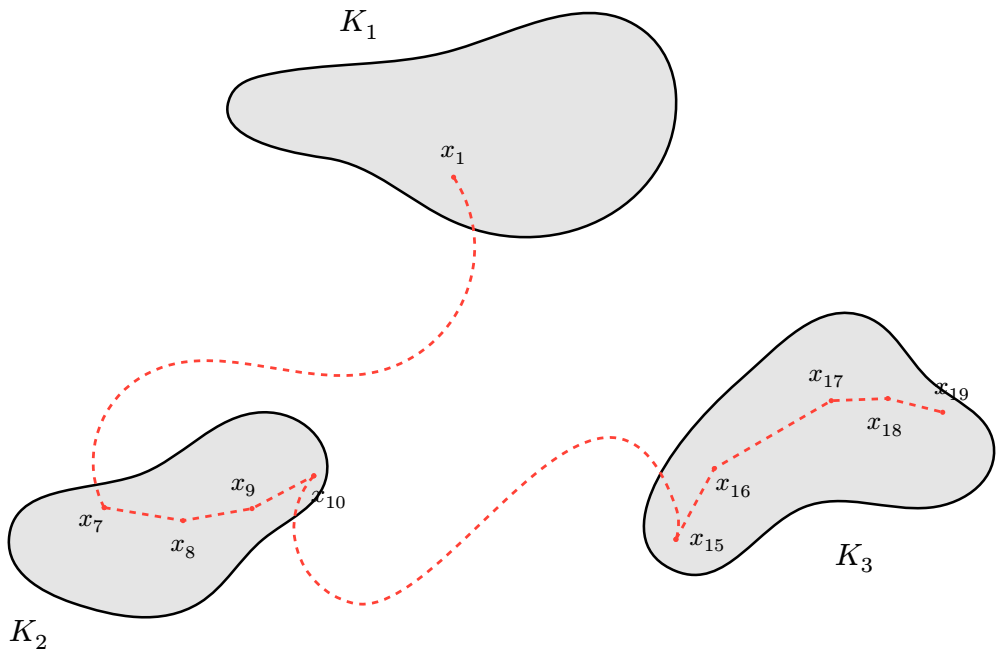
Main idea of the proof

→ Restrict SGD to a chain visiting only critical components



Main idea of the proof

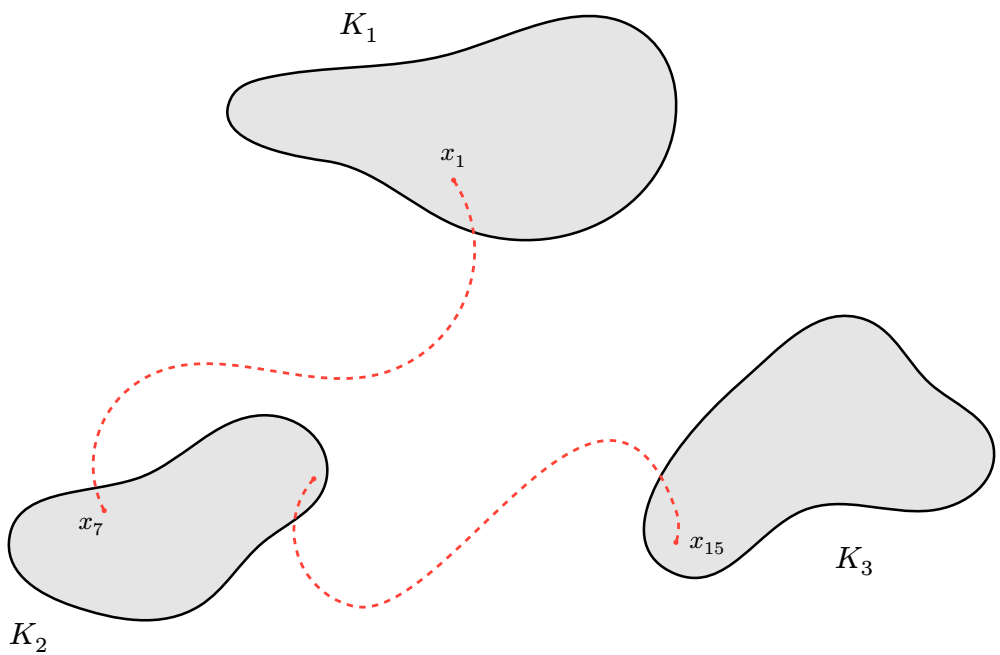
→ Restrict SGD to a chain visiting only critical components



Forget what happens outside: negligeable time

Main idea of the proof

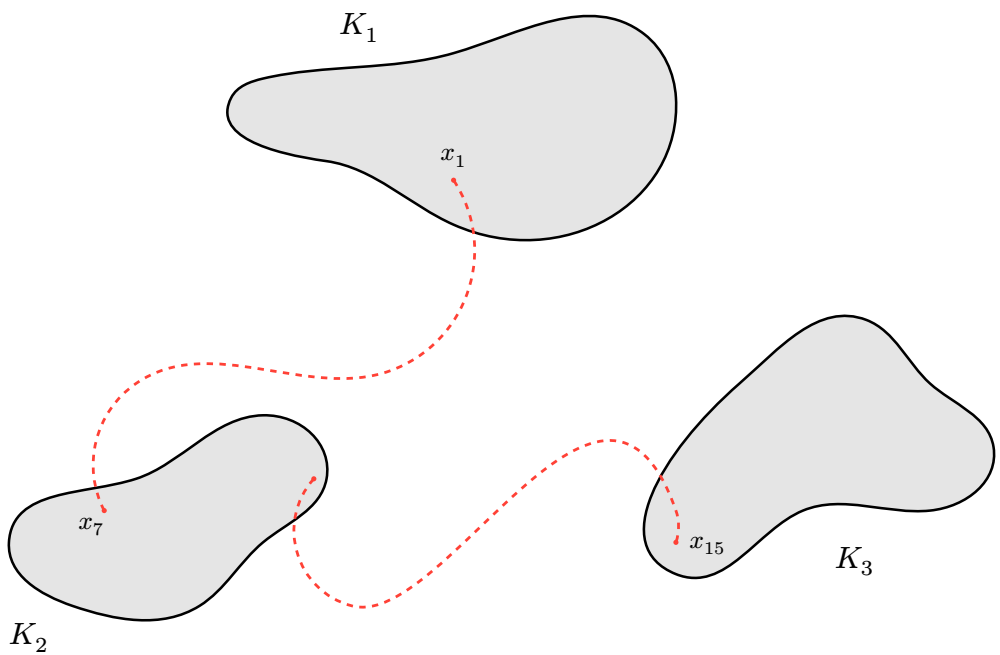
→ Restrict SGD to a chain visiting only critical components



Forget what happens outside: negligeable time
Abstract away what happens inside: does not depend on the starting point

Main idea of the proof

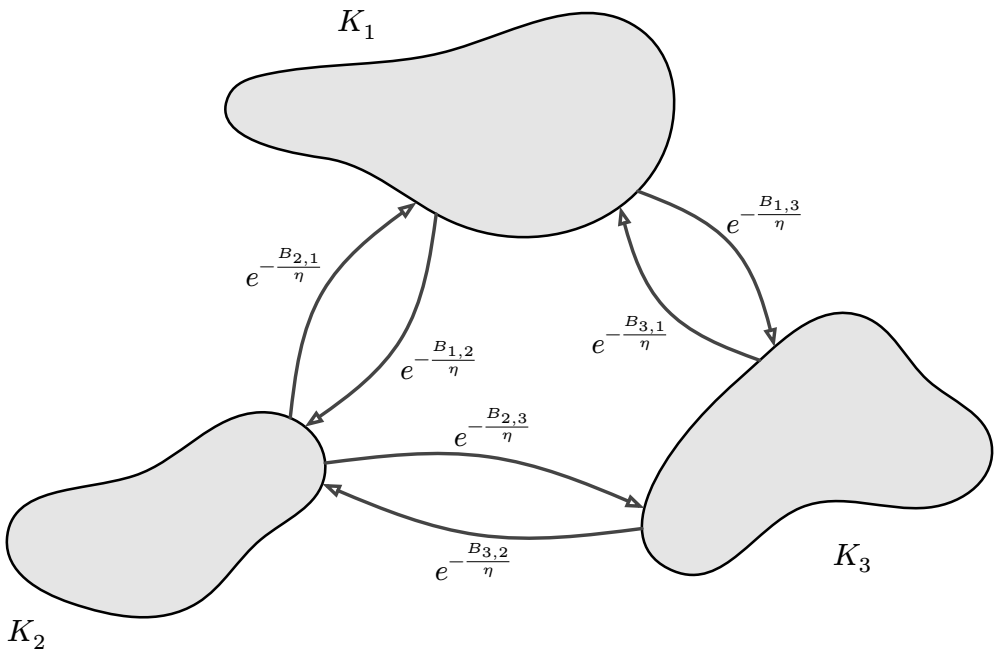
→ Restrict SGD to a chain visiting only critical components



Forget what happens outside: negligeable time
Abstract away what happens inside: does not depend on the starting point
Keep only transitions

Main idea of the proof

$\rightarrow$ Restrict SGD to a chain visiting only critical components



Forget what happens outside: negligeable time
Abstract away what happens inside: does not depend on the starting point
Keep only transitions
$\rightarrow$ Study SGD as a finite-state space Markov chain on $\{K_1, ..., K_p\}$ with transition probabilities

$$p_{i,j} \sim e^{-\frac{B_{i,j}}{\eta}}$$

Energy

Using exact formulas for finite-state space Markov chains:

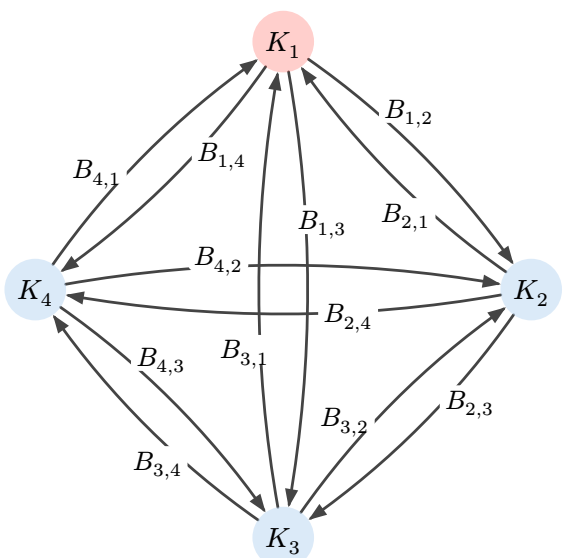
Lemma *(very informal)**:** the invariant measure of SGD restricted to* $\{K_1, ..., K_p\}$ *is, for* $\eta > 0$ *small enough,*

$$\pi(i) \propto \exp\left(-\frac{E_i}{\eta}\right)$$

where ***energy*** *of* K_i*:*

$$E_i = \min\left\{ \sum_{j \to k \in T} B_{j,k} \mid T \text{ spanning tree pointing to } i \right\}$$

$E_i =$ "min. cost to join all components to i"



Energy

Using exact formulas for finite-state space Markov chains:

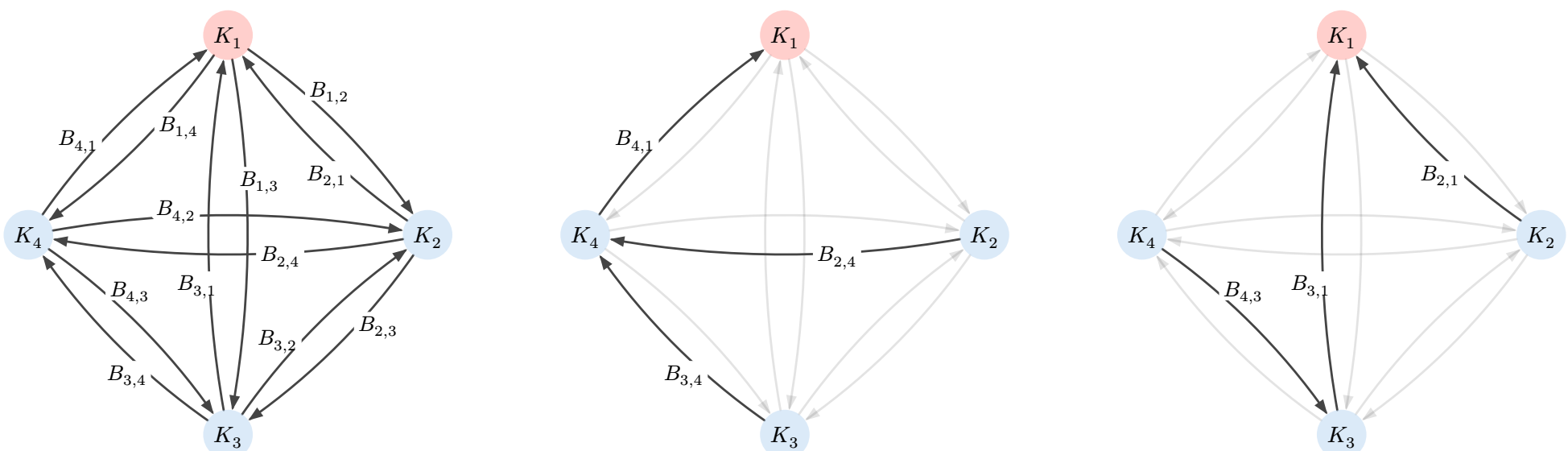
Lemma *(very informal)**:** the invariant measure of SGD restricted to* $\{K_1, ..., K_p\}$ *is, for* $\eta > 0$ *small enough,*

$$\pi(i) \propto \exp\left(-\frac{E_i}{\eta}\right)$$

where ***energy*** *of* K_i*:*

$$E_i = \min\left\{ \sum_{j \to k \in T} B_{j,k} \mid T \text{ spanning tree pointing to } i \right\}$$

$E_i =$ "min. cost to join all components to i"



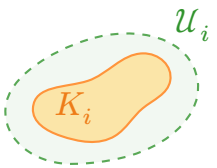
On the long-run behavior of SGD in non-convex landscapes

1. **Introduction** Context and informal overview of Result 1
2. **Our Approach** Essential ideas, key objects and proof sketch of Result 1
3. **Result 1** Long-run distribution of SGD
4. **Result 2** Time to reach the global minimum
5. **Perspectives** Beyond SGD

Result 1: Invariant measure of SGD

Theorem: *Consider μ_∞ any invariant measure of SGD:*

Given $\varepsilon > 0$, $\mathcal{U}_i$ neighborhoods of K_i, and $\eta > 0$ small enough:



1. ***Concentration near critical points:***

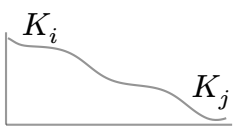
$$\mu_\infty\left(\bigcup_{i=1}^{p} \mathcal{U}_i\right) \geq 1 - e^{-\frac{c}{\eta}}, \qquad \text{for some } c > 0$$

2. ***Boltzmann-Gibbs distribution:*** *for all i,*

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

3. ***Saddle-point avoidance:*** *if K_i is a saddle, then there is K_j local minimum with $E_j < E_i$:*

$$\frac{\mu_\infty(\mathcal{U}_i)}{\mu_\infty(\mathcal{U}_j)} \leq e^{-\frac{c}{\eta}} \qquad \text{for some } c > 0$$



4. ***Ground state concentration:*** *given $\mathcal{U}_0$ neighborhood of the ground states $K_0 = \operatorname{argmin}_i E_i$*

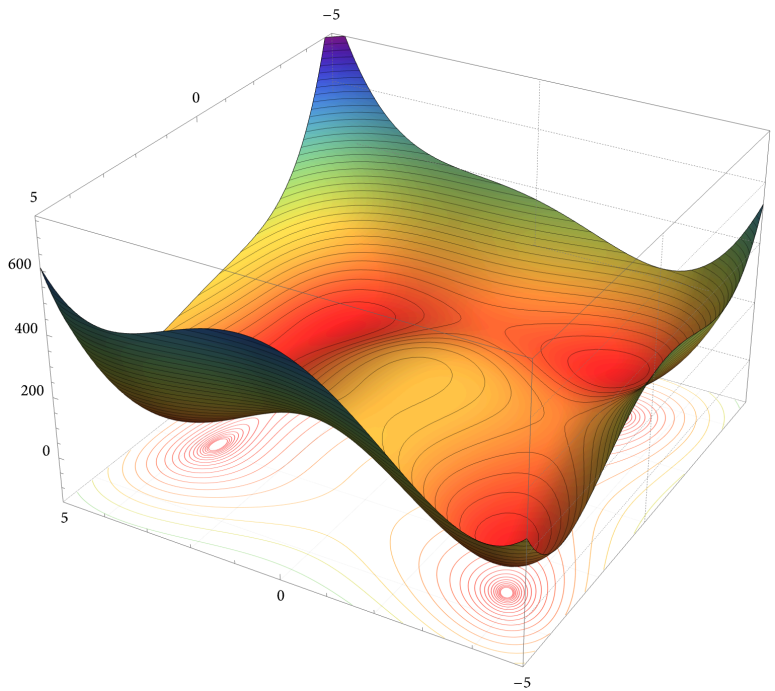
$$\mu_\infty(\mathcal{U}_0) \geq 1 - e^{-\frac{c}{\eta}}, \qquad \text{for some } c > 0$$

Example: Gaussian noise

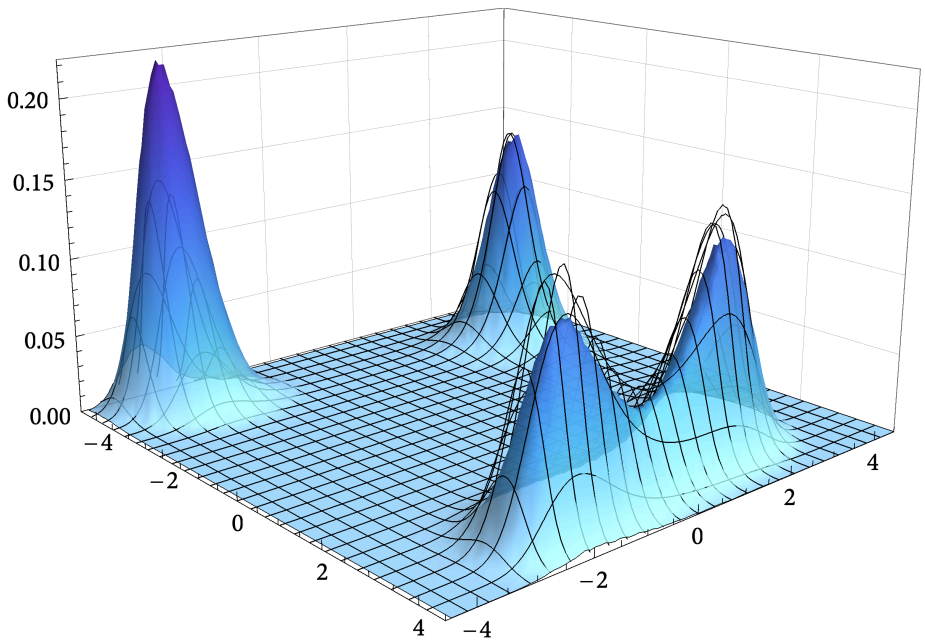
Assume $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I_d)$

Boltzmann-Gibbs distribution: for all i,

$$E_i = \frac{2f(K_i)}{\sigma^2} \quad \Rightarrow \quad \mu_\infty(\mathcal{U}_i) \approx \exp\left(-\frac{2f(K_i)}{\sigma^2\eta}\right) \quad \text{and} \quad K_0 = \operatorname{argmin} f$$

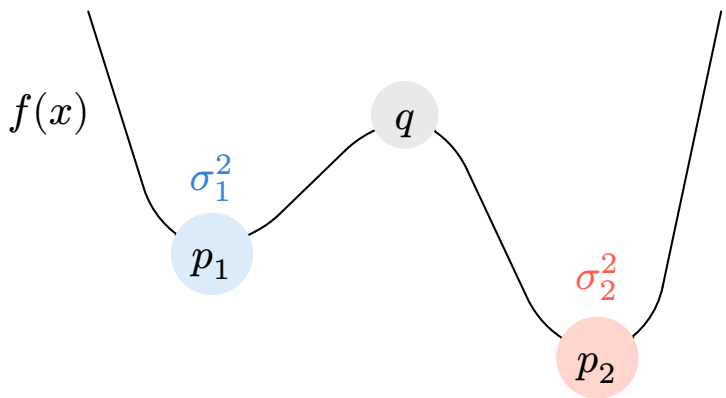


Himmelblau function

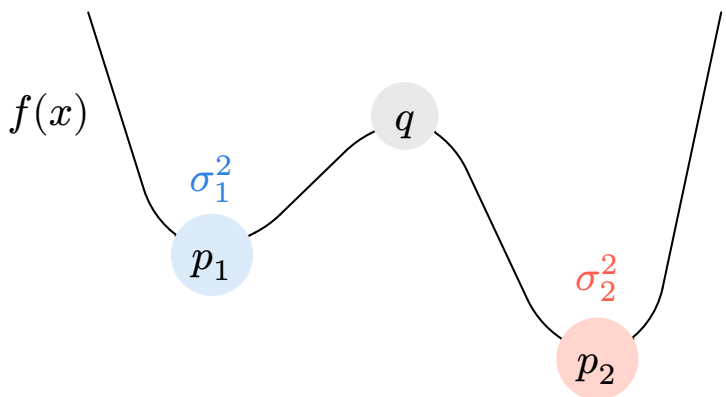


Simulation (shaded blue) vs prediction (black wireframe) of the invariant measure

Minimizers of the energy = minimizers of the function?

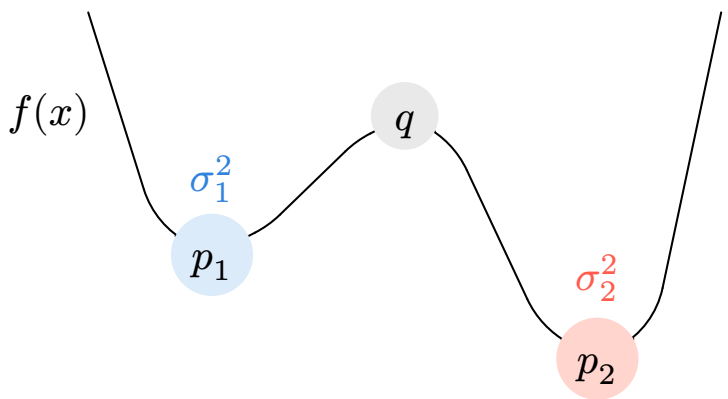


Minimizers of the energy = minimizers of the function?



$$E_1 = \frac{f(q) - f(p_2)}{\sigma_2^2} \qquad \text{and} \qquad E_2 = \frac{f(q) - f(p_1)}{\sigma_1^2}$$

Minimizers of the energy = minimizers of the function?



$$E_1 = \frac{f(q) - f(p_2)}{\sigma_2^2} \qquad \text{and} \qquad E_2 = \frac{f(q) - f(p_1)}{\sigma_1^2}$$

$$\sigma_1 \text{ small enough} \quad \Rightarrow \quad E_1 < E_2 \quad \Rightarrow \quad \mu_\infty \text{ concentrates on } p_1$$

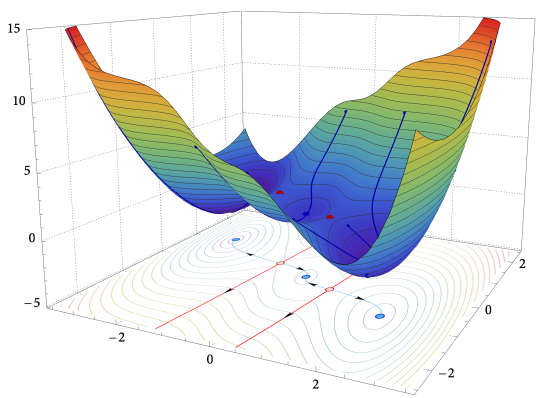
$\rightarrow$ In general, minimizer of the energy $\neq$ minimizer of the function!

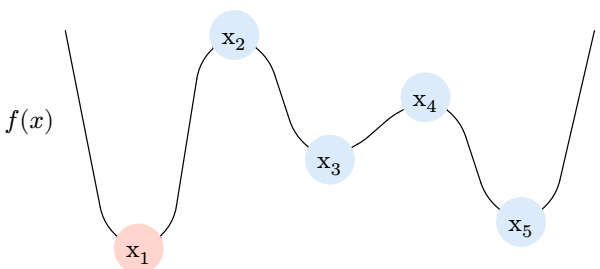
On the long-run behavior of SGD in non-convex landscapes

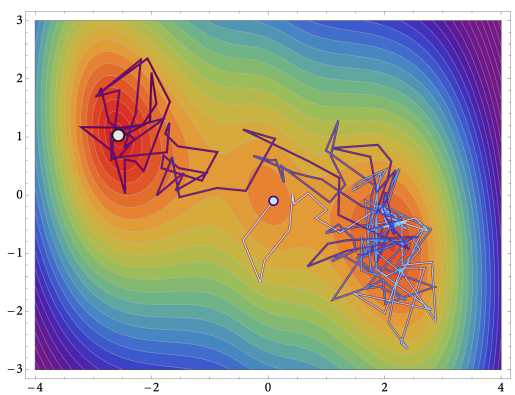
1. **Introduction** Context and informal overview of Result 1
2. **Our Approach** Essential ideas, key objects and proof sketch of Result 1
3. **Result 1** Long-run distribution of SGD
4. **Result 2** Time to reach the global minimum
5. **Perspectives** Beyond SGD

Time to reach the global minimum

Q2: How much time does SGD take to reach the global minimum?







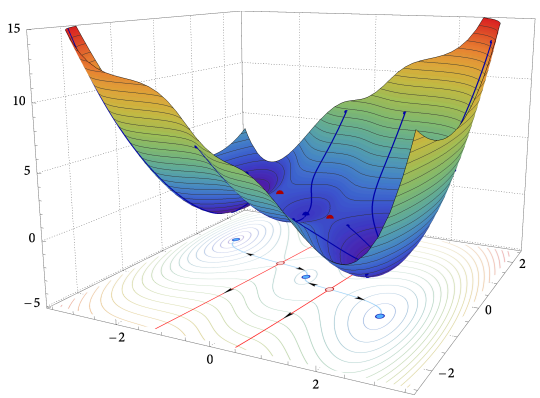
Time to reach the global minimum

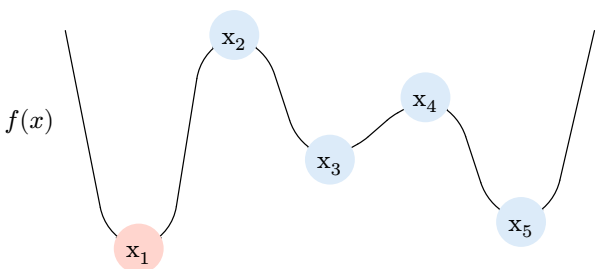
Q2: How much time does SGD take to reach the global minimum?

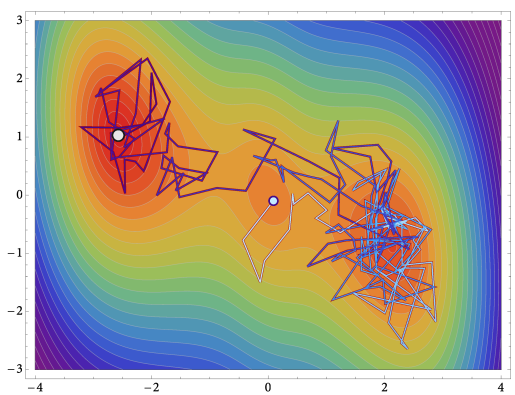
Hitting time: with small margin $\delta > 0$,

$$\tau = \min\{t \in \mathbb{N} \mid \operatorname{dist}(x_t, \operatorname{argmin} f) \leq \delta\}$$

Q2: What is $\mathbb{E}_x[\tau]$ for SGD started at x?



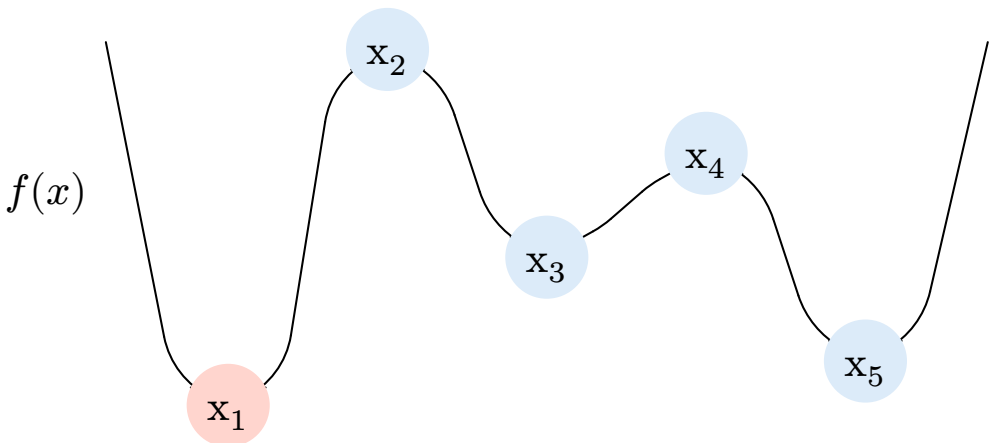




Example: Three Humped Camel

$Z(x;\omega) \sim N(0, \sigma^2 I_d)$

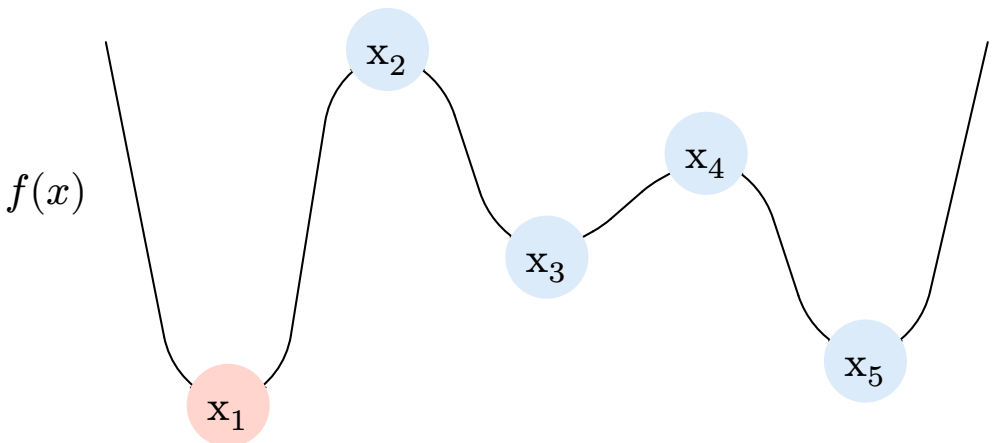
Q: What is $\mathbb{E}_x[\tau]$ for SGD started at x_3 here?



Example: Three Humped Camel

$Z(x;\omega) \sim N(0, \sigma^2 I_d)$

Q: What is $\mathbb{E}_x[\tau]$ for SGD started at x_3 here?



(Informal) Theorem:

$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{f(x_2) - f(x_5)}{\sigma^2 \eta}\right)$$

Result 2: Time to reach the global minimum

Theorem:
starting at x, the time τ to reach $\operatorname{argmin} f$ satisfies, for any $\varepsilon > 0$ and $\eta, \delta > 0$ small enough,

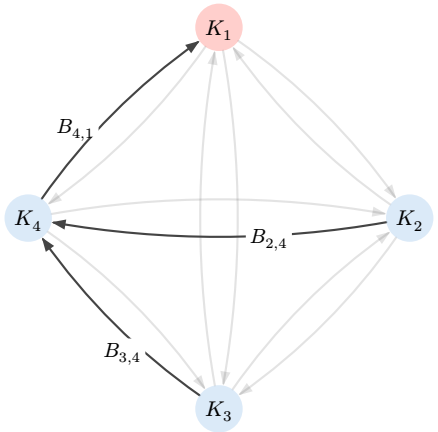
$$\exp\left(\frac{J(x) - \varepsilon}{\eta}\right) \;\leq\; \mathbb{E}_x[\tau] \;\leq\; \exp\left(\frac{J(x) + \varepsilon}{\eta}\right)$$

provided a technical condition holds for the LHS.

Key quantity $J(x)$: geometric measure of problem's hardness, it captures

- The difficulty of the loss landscape: hardest set of obstacles to overcome to reach $\operatorname{argmin} f$
- The statistics of the noise: scales with inverse square of the noise level

Definition of $J(x)$



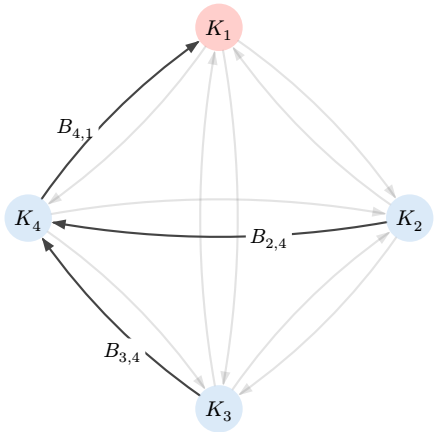
Energy of $K_1 = \operatorname{argmin} f$: *"min. cost to join everyone to argmin"*

$$E_1 = \min\left\{ \sum_{j \to k \in T} B_{j,k} \mid T \text{ spanning tree pointing to } 1 \right\}$$

Energy of pruning K_i: *"min. cost to join everyone to argmin except i"*

$$J(i \nrightarrow 1) = \min\left\{ \sum_{j \to k \in T} B_{j,k} \mid T \text{ spanning tree pointing to } 1 \text{ with an edge from } i \text{ to } 1 \text{ removed} \right\}$$

Definition of $J(x)$



Energy of $K_1 = \operatorname{argmin} f$: *"min. cost to join everyone to argmin"*

$$E_1 = \min\left\{ \sum_{j \to k \in T} B_{j,k} \mid T \text{ spanning tree pointing to } 1 \right\}$$

Energy of pruning K_i: *"min. cost to join everyone to argmin except i"*

$$J(i \nrightarrow 1) = \min\left\{ \sum_{j \to k \in T} B_{j,k} \mid T \text{ spanning tree pointing to } 1 \text{ with an edge from } i \text{ to } 1 \text{ removed} \right\}$$

Energy of $\operatorname{argmin} f$ relative to K_i: *"cost of connecting i to argmin = cost of obstacles from i to argmin"*

$$J(i) = E_1 - J(i \nrightarrow 1)$$

Energy of $\operatorname{argmin}(f)$ relative to x: *"cost of all possible obstacles from x to argmin"*

$$J(x) = \max_{i=2,\dots,p} [J(i) - B(x,i)]_+$$

Core Takeaway

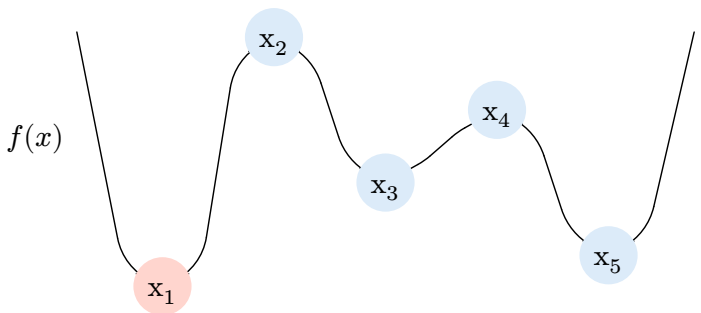
Assume $Z(x;\omega)$ has noise strength between $\sigma_{\min}$ and $\sigma_{\max}$ in every direction.

Theorem: *for any starting point* x

$$J(x) \quad \leq \quad 2\frac{\max_{K_i \neq \operatorname{argmin} f} f(K_i) - \min_{K_i \neq \operatorname{argmin} f} f(K_i)}{\sigma_{\min}^2} \quad + \quad \mathcal{O}\left(1 - \frac{\sigma_{\min}^2}{\sigma_{\max}^2}\right)$$

$$J(x) \leq \frac{2 \times (f(x_2) - f(x_5))}{\sigma^2}$$

$$\mathbb{E}_x[\tau] \leq \exp\left(\frac{f(x_2) - f(x_5)}{\sigma^2 \eta}\right)$$



Core Takeaway

Assume $Z(x;\omega)$ has noise strength between $\sigma_{\min}$ and $\sigma_{\max}$ in every direction.

Theorem: *for any starting point* x

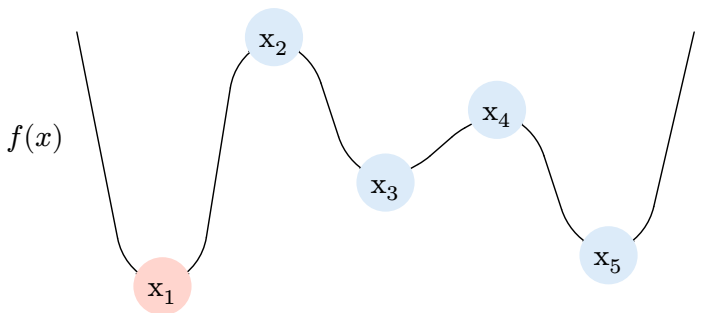
$$J(x) \quad \leq \quad 2\frac{\max_{K_i \neq \operatorname{argmin} f} f(K_i) - \min_{K_i \neq \operatorname{argmin} f} f(K_i)}{\sigma_{\min}^2} \quad + \quad \mathcal{O}\left(1 - \frac{\sigma_{\min}^2}{\sigma_{\max}^2}\right)$$

$\rightarrow$ Tight!

For x close to x_3,

$$J(x) = \frac{2 \times (f(x_2) - f(x_5))}{\sigma^2}$$

$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{f(x_2) - f(x_5)}{\sigma^2 \eta}\right)$$



Core Takeaway

Assume $Z(x;\omega)$ has noise strength between $\sigma_{\min}$ and $\sigma_{\max}$ in every direction.

Theorem: *for any starting point* x

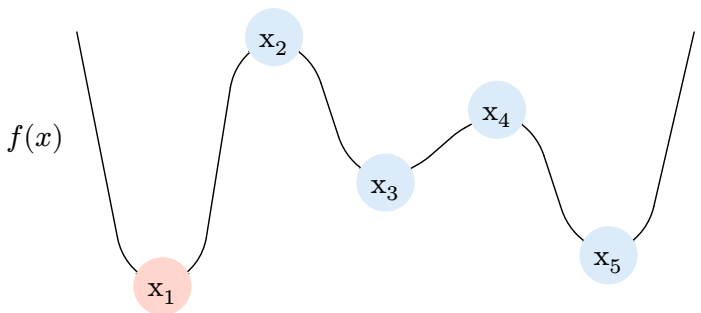
$$J(x) \quad \leq \quad 2\frac{\max_{K_i \neq \operatorname{argmin} f} f(K_i) - \min_{K_i \neq \operatorname{argmin} f} f(K_i)}{\sigma_{\min}^2} \quad + \quad \mathcal{O}\left(1 - \frac{\sigma_{\min}^2}{\sigma_{\max}^2}\right)$$

$\rightarrow$ Tight!

For x close to x_3,

$$J(x) = \frac{2 \times (f(x_2) - f(x_5))}{\sigma^2}$$

$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{f(x_2) - f(x_5)}{\sigma^2 \eta}\right)$$



$\rightarrow$ Non-convex optimization is “easy” when

depths of spurious minima $\ll$ $\eta \times$ local noise strength

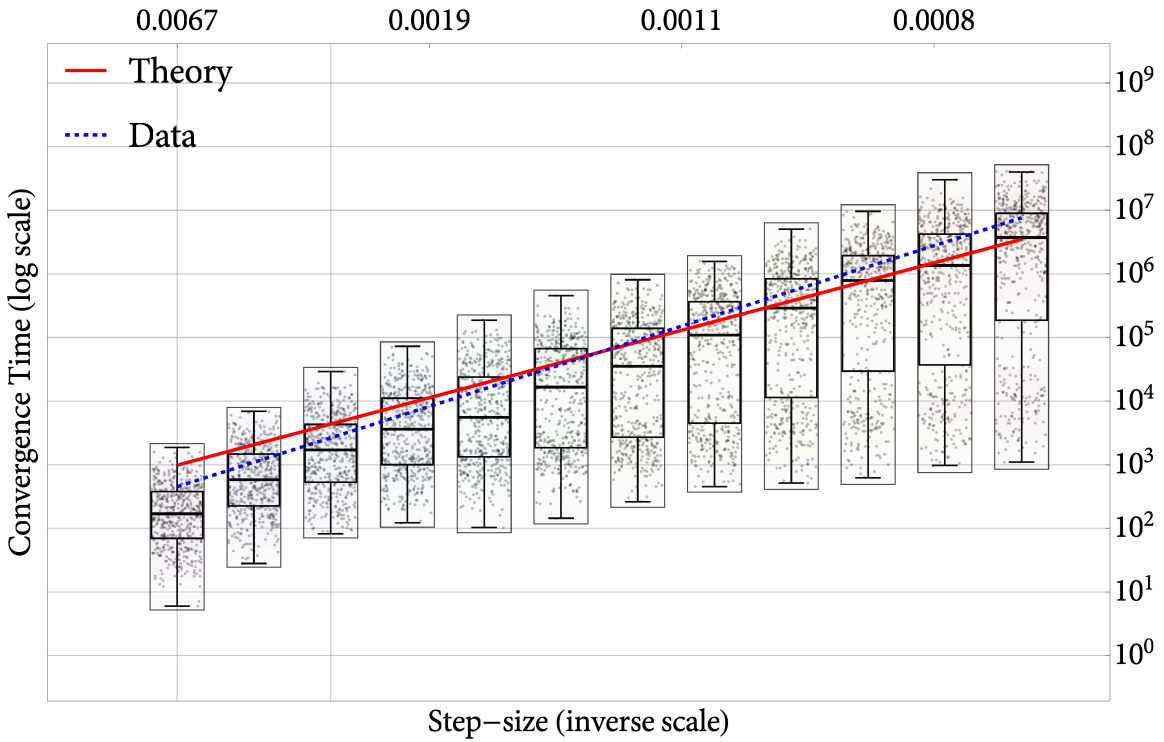
Three Humps: Simulation

For $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I_d)$,

we predict

$$J(x) = \frac{2(f(x_2) - f(x_5))}{\sigma^2}$$

$$\log \mathbb{E}[\tau] \approx J(x) \times \frac{1}{\eta}$$



Linear fit to data: $R^2 = 0.99$; Theoretical prediction: $R^2 = 0.97$

Non-constant noise?

Assume $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 f(x) I_d)$

$\rightarrow$ Relevant for deep learning, eg (Mori et al., 2022)

Theorem: *for any starting point* x

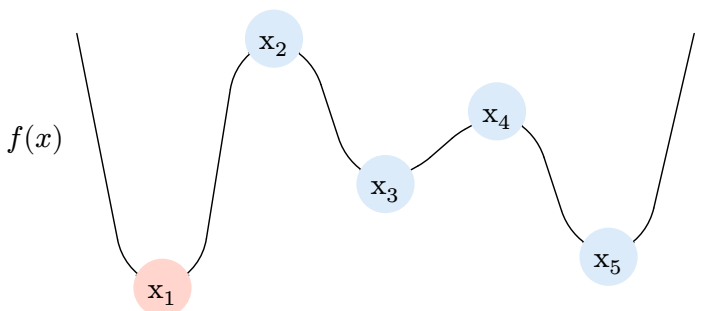
$$J(x) \leq 2\frac{\max_{K_i \neq \operatorname{argmin} f} \log f(K_i) - \min_{K_i \neq \operatorname{argmin} f} \log f(K_i)}{\sigma^2}$$

$\rightarrow$ Tight!

For x close to x_3,

$$J(x) = \frac{2 \times (\log f(x_2) - \log f(x_5))}{\sigma^2}$$

$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{\log f(x_2) - \log f(x_5)}{\sigma^2 \eta}\right)$$



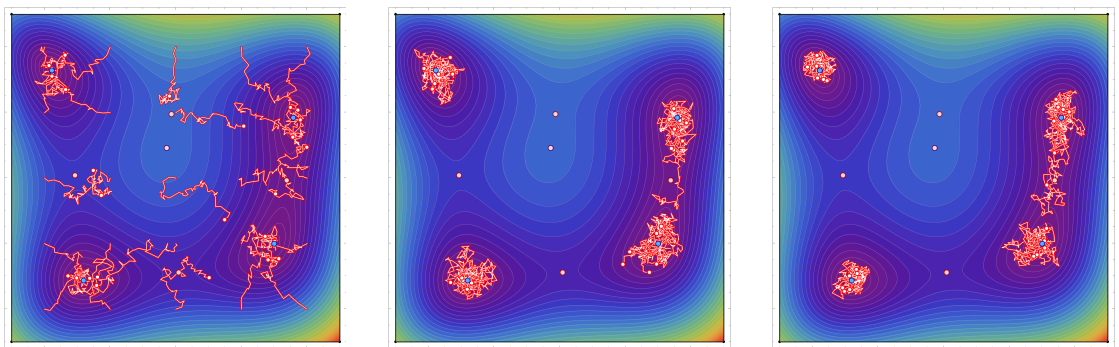
$\rightarrow$ Non-convex optimization is "easy" when

log depths of spurious minima $\ll$ $\eta \times$ local noise strength

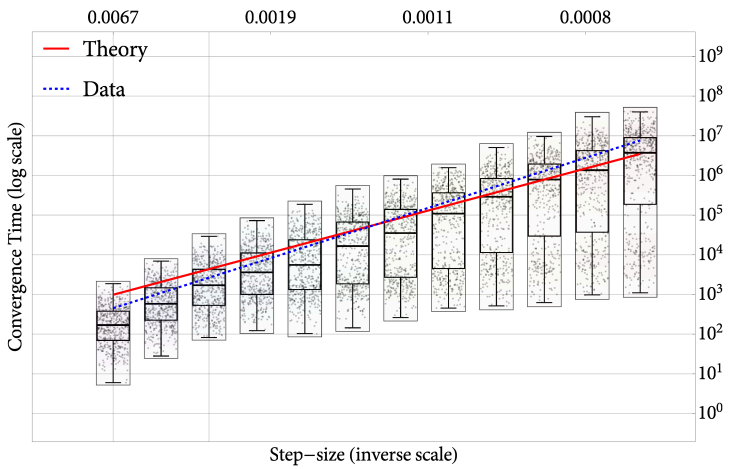
Main takeaways

Constant step size SGD on non-convex functions:

- Iterates does not converge but distribution can be studied
- It concentrates near local minima and in particular near those that minimize the energy
- It takes exponential time to reach the global minima, with exponent depending on the geometry of the loss and the noise



$t = 50$ $t = 200$ $t = 500$



On the long-run behavior of SGD in non-convex landscapes

1. **Introduction** Context and informal overview of Result 1
2. **Our Approach** Essential ideas, key objects and proof sketch of Result 1
3. **Result 1** Long-run distribution of SGD
4. **Result 2** Time to reach the global minimum
5. **Perspectives** Beyond SGD

Project 1: Understanding adaptive methods

In practice, AdaGrad / Adam often converge faster than SGD.

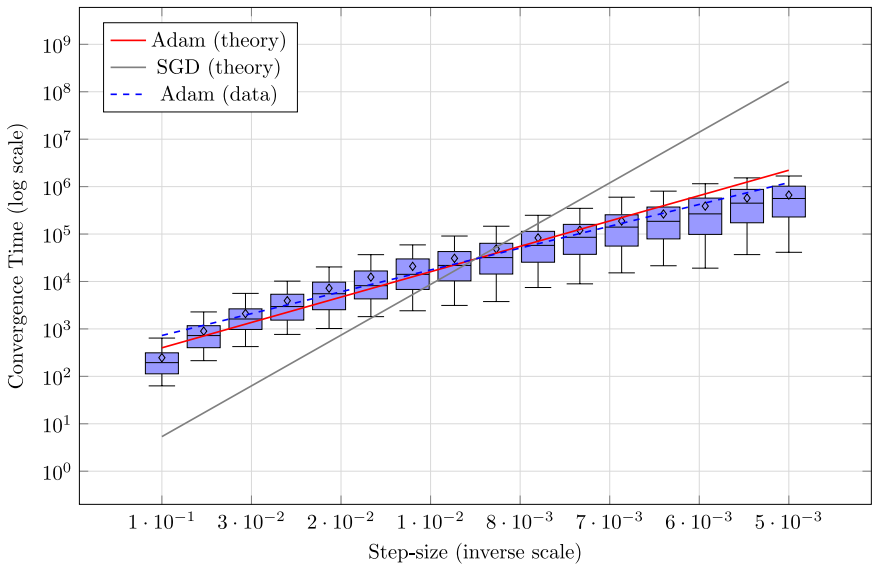
Q1: Why do adaptive methods often outperform SGD?

→ Need to adapt our framework, which relies heavily on the gradient structure of SGD.

Preliminary results:

in some simple cases,

$$\mathbb{E}[\tau_{\text{Adam}}] \approx \sqrt{\mathbb{E}[\tau_{\text{SGD}}]}$$



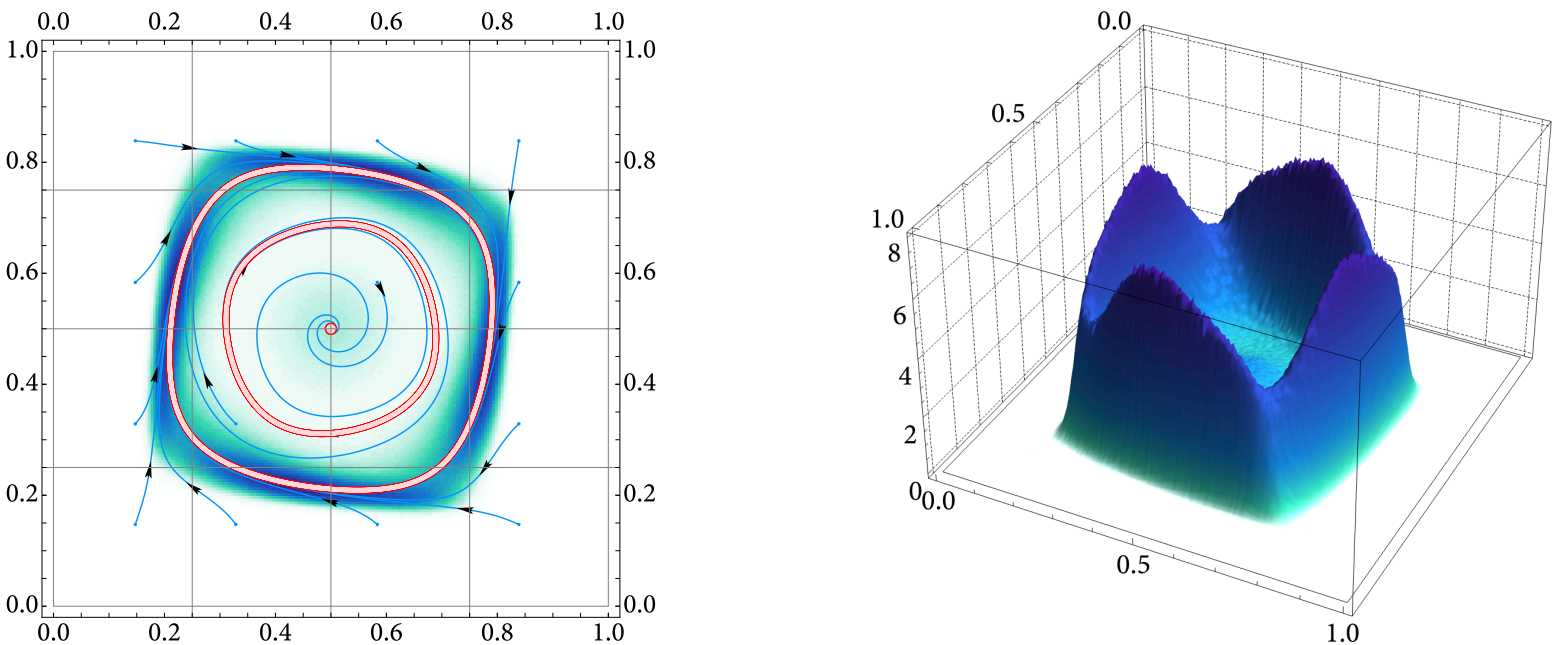
Hitting time of Adam in a non-convex problem

Project 2: Characterizing learning dynamics in games

In many settings, learning emerges from interactions between multiple agents (e.g. GANs, multi-agent RL)

Q2: What is the long-run behavior of learning dynamics in general games?

→ Solution sets are not only fixed points but also cycles and more complex attractors



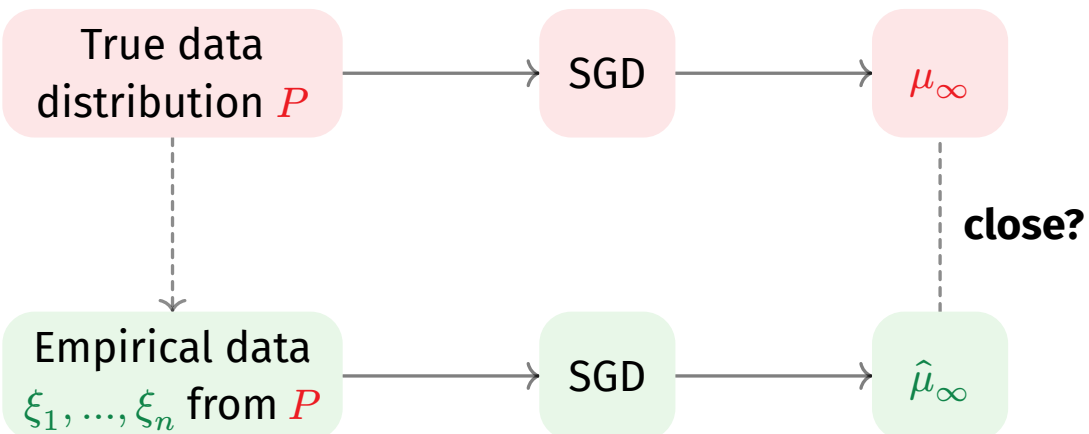
Asymptotic distribution of players' strategies in a non-concave game

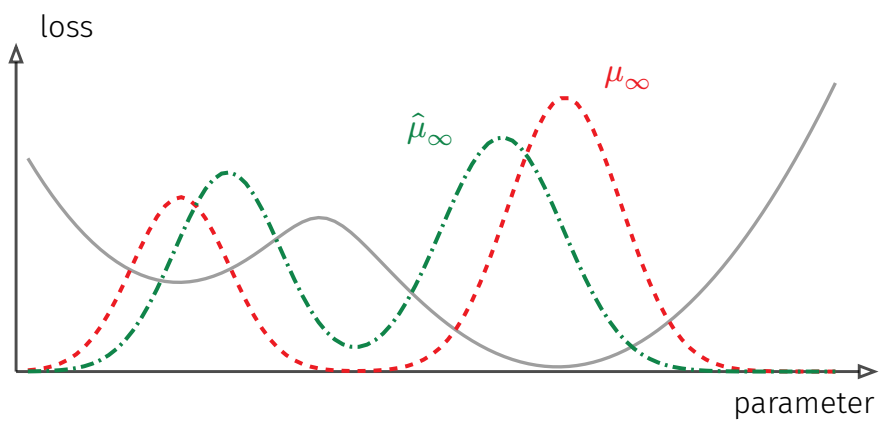
Project 3: From training dynamics to generalization

Understanding why neural networks generalize well remains a fundamental question in deep learning theory.

Q3: How do training dynamics shape the generalization performance of the end model?

→ **Possible Direction:** study the stability of the invariant measure of SGD.





Towards a principled understanding of stochastic optimization in deep learning

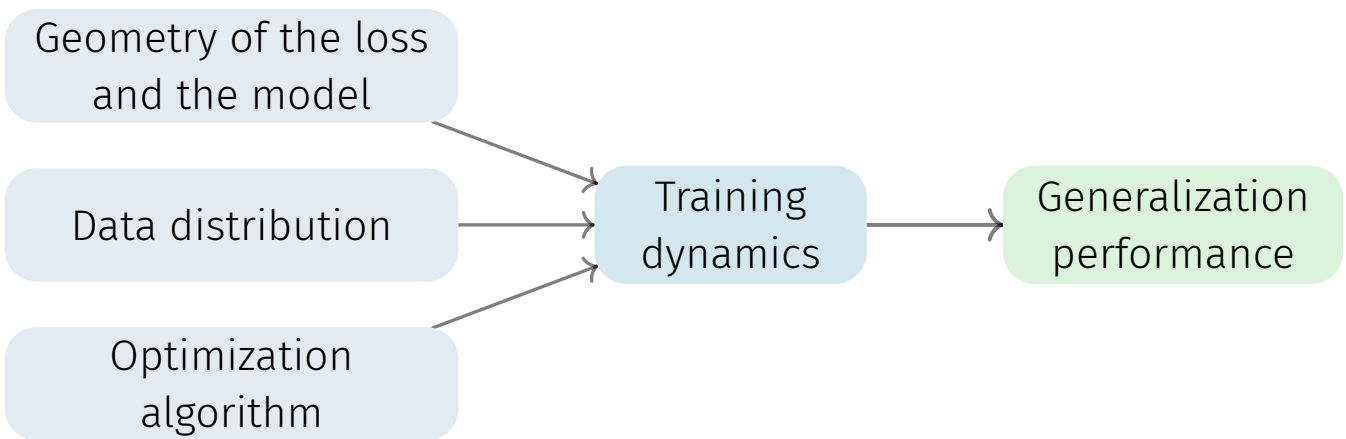
Short-term goals:

Project 1: Understanding adaptive methods (Adam, AdaGrad)

Project 2: Characterizing learning dynamics in games (GANs, multi-agent RL)

Project 3: From training dynamics to generalization

Long-term goal: *A principled understanding of non-convex optimization for statistical learning*



Publications

Publications mentionned today

What is the long-run distribution of stochastic gradient descent? A large deviations analysis.
ICML, 2024.

Regularization for Wasserstein distributionally robust optimization.
ESAIM: COCV, 2023.

The rate of convergence of Bregman proximal methods: Local geometry vs. regularity vs. sharpness.
SIAM Journal on Optimization, 2024.

The global convergence time of stochastic gradient descent in non-convex landscapes.
ICML, 2025.

Exact generalization guarantees for (regularized) Wasserstein distributionally robust models.
NeurIPS, 2023.

The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities.
COLT, 2021.

Other publications

Skwdro: A library for Wasserstein distributionally robust machine learning.
JMLR, 2026.

Almost sure convergence of stochastic gradient methods under gradient domination.
TMLR, 2025.

Expressive power of invariant and equivariant graph neural networks.
ICLR, 2021.

Linear lower bounds and conditioning of differentiable games.
ICML, 2020.

How does the pretraining distribution shape in-context learning? A fundamental trade-off
ICML, 2026.

The geometries of truth are orthogonal across tasks.
ICML Workshop, 2025.

A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games.
AISTATS, 2020.

Accelerating smooth games by manipulating spectral shapes.
AISTATS, 2020.

Full texts and references: wazizian.fr