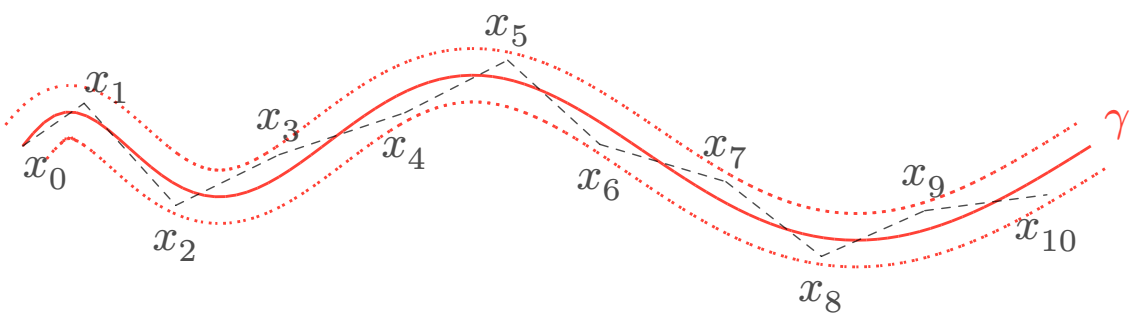
The Long-Run Behavior of SGD in Non-Convex Landscapes

A Large Deviation Analysis



SMAI MODE 2026, Nice, France

March 16, 2026

W. Azizian, F. Iutzeler, J. Malick, P. Mertikopoulos

Deep learning = nonconvex loss landscape

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

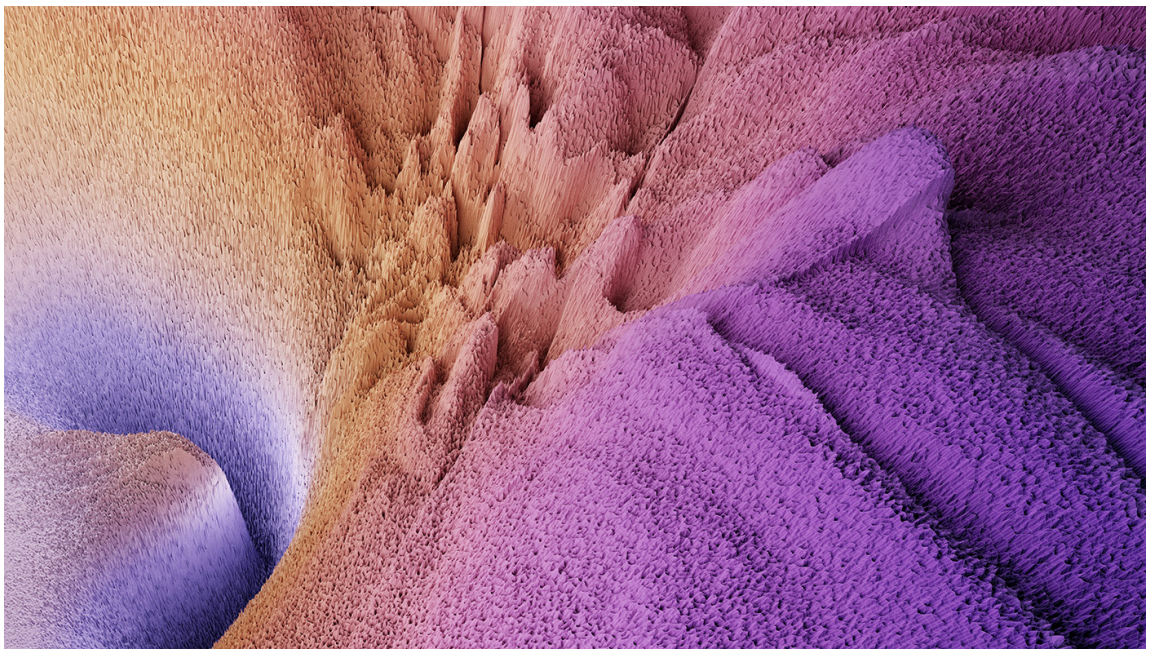


Image credit: losslandscape.com

Deep learning = nonconvex loss landscape

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

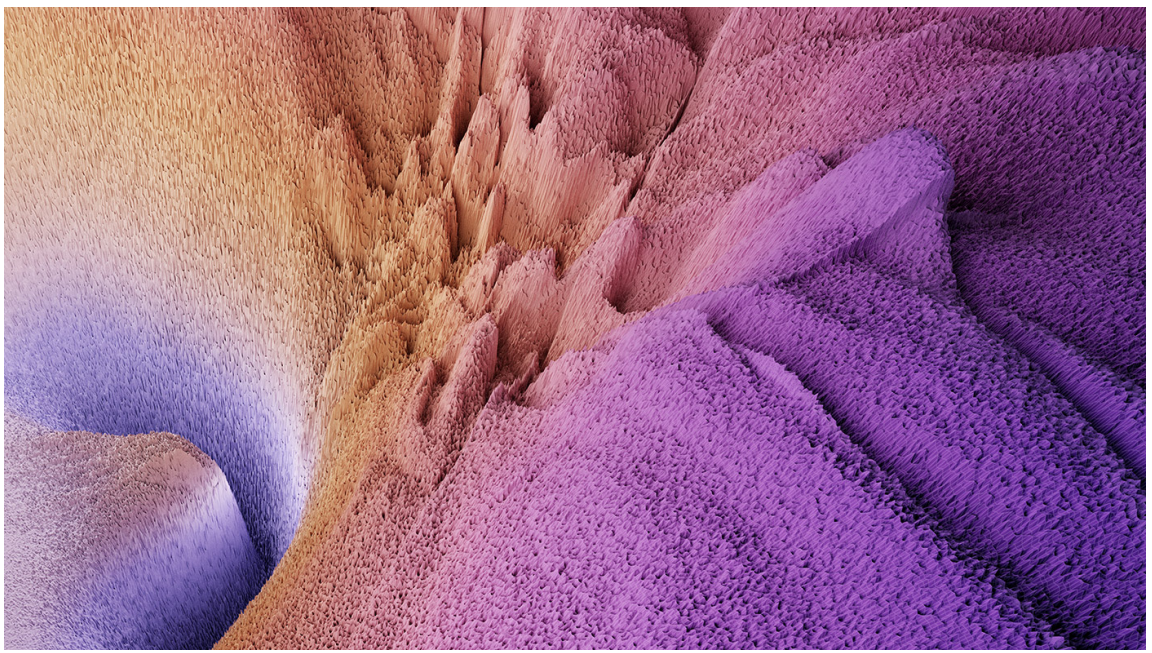


Image credit: losslandscape.com

Q: What is the long-run behavior of stochastic gradient methods?

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 $\eta > 0$

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

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,

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

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 $\eta > 0$

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

What is known when f is non-convex:

- 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)

Stochastic Gradient Descent

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

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

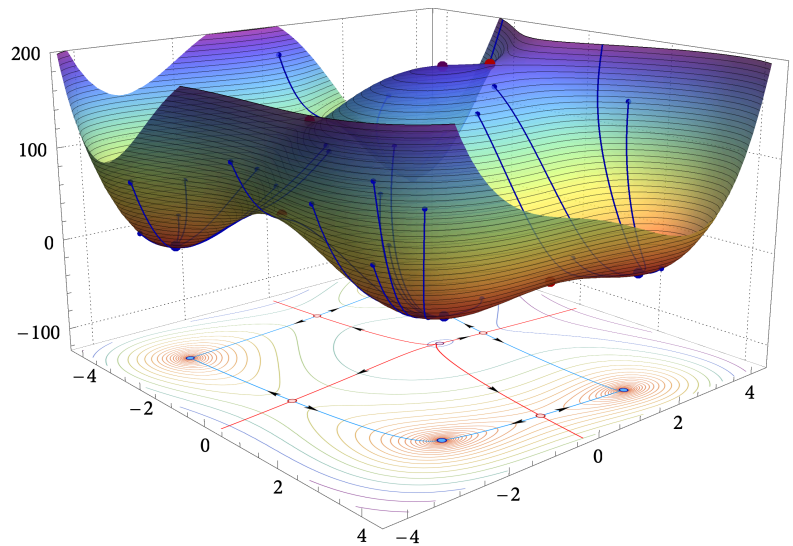
Stochastic Gradient Descent (SGD): with *constant* step-size $\eta > 0$

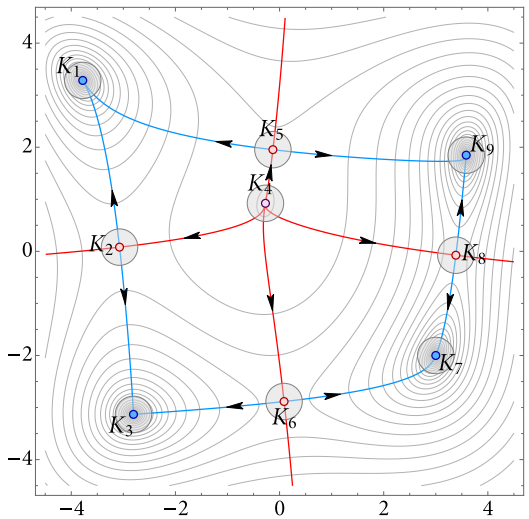
$$x_{t+1} = x_t - \underbrace{\eta}_{\text{step-size}} \Big[\nabla f(x_t) + \underbrace{Z(x_t; \omega_t)}_{\text{zero-mean noise}} \Big]$$

Q: What is the asymptotic behavior of SGD?

Example: Himmelblau function

Himmelblau function



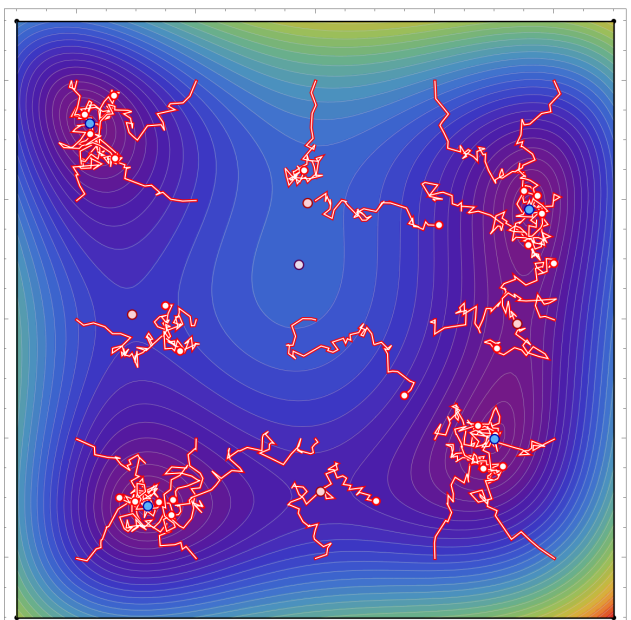


In our work, critical set of f:

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

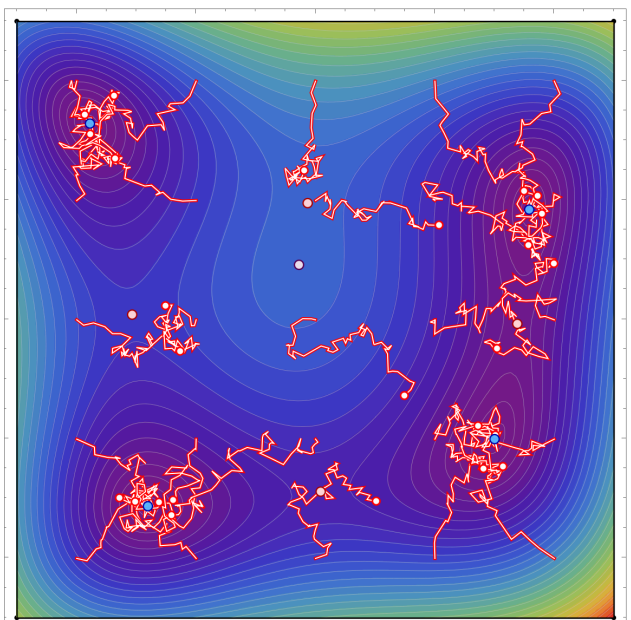
where K_i connected components (compact)

Example: SGD on Himmelblau function

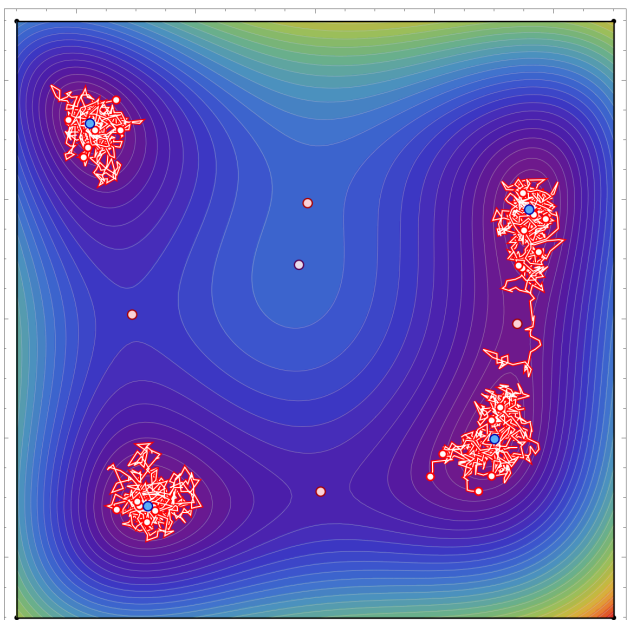


$t = 50$

Example: SGD on Himmelblau function

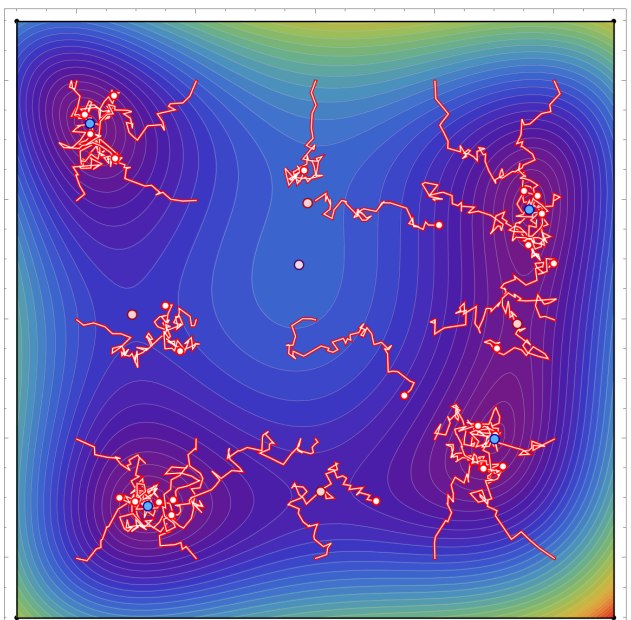


$t = 50$

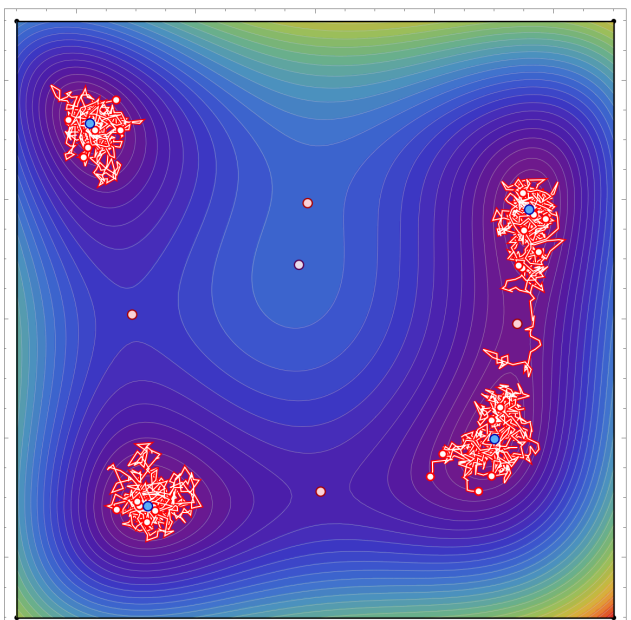


$t = 200$

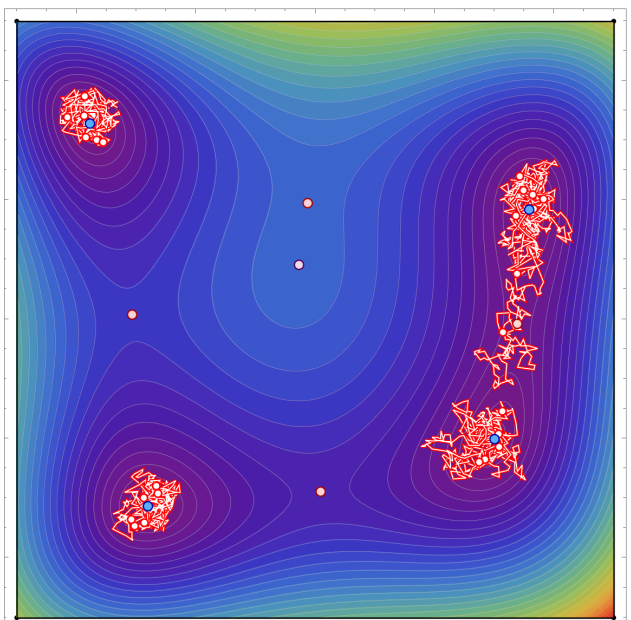
Example: SGD on Himmelblau function



$t = 50$



$t = 200$



$t = 500$

Asymptotic distribution of SGD

SGD with *constant* step-size = Markov Chain

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

Invariant measure: probability measure μ_∞ such that

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

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 $n \to \infty$,

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

Q: Where does the invariant measure of SGD concentrate?

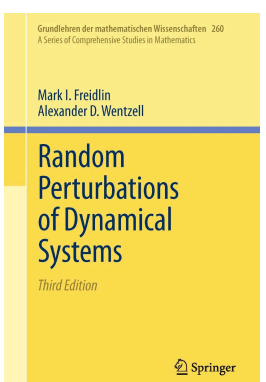
New Approach: large deviations

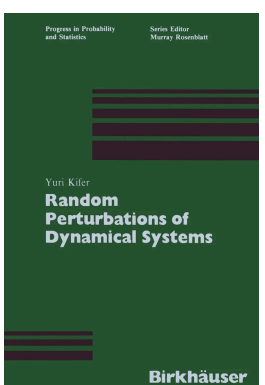
This work: we characterize where μ_∞ is concentrated through a large deviation approach

- Based on large deviation theory and the theory of random perturbations of dynamical systems
- We adapt & refine this theory for SGD (lack of compactness, realistic noise models, discrete-time)

In this talk:

- Introduce the essential ideas of this approach
- Present our main results





Based on our paper: *What is the long-run behavior of SGD? A large deviation analysis.* ICML 2024

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 connected components:

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

Noise assumptions:

- $\mathbb{E}[Z(x;\omega)] = 0$, $\mathrm{cov}(Z(x;\omega)) \succ 0$, $Z(x;\omega) = O(\|x\|)$ 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 connected components:

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

Noise assumptions:

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

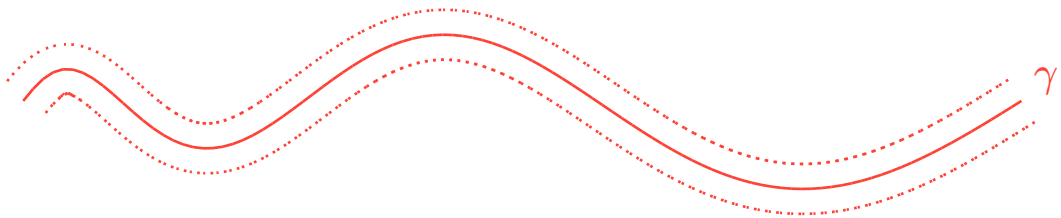
$$\log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right] \leq \frac{\sigma^2}{2}\|v\|^2$$

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.

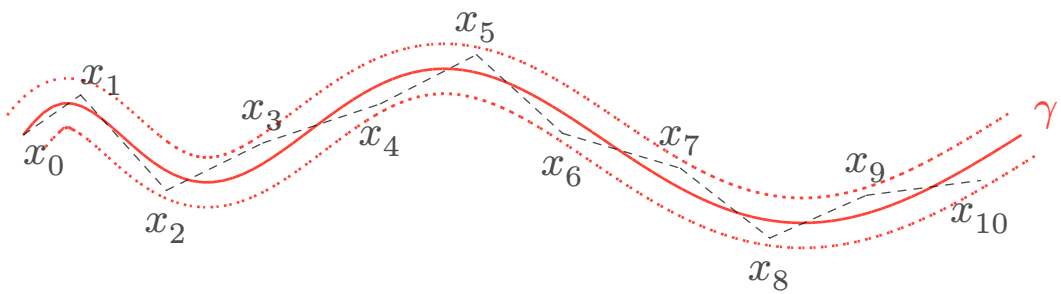
Large deviations for discrete-time SGD

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



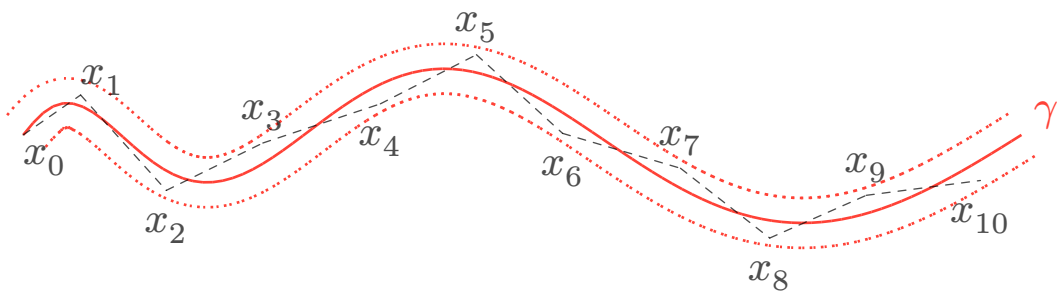
Large deviations for discrete-time SGD

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)

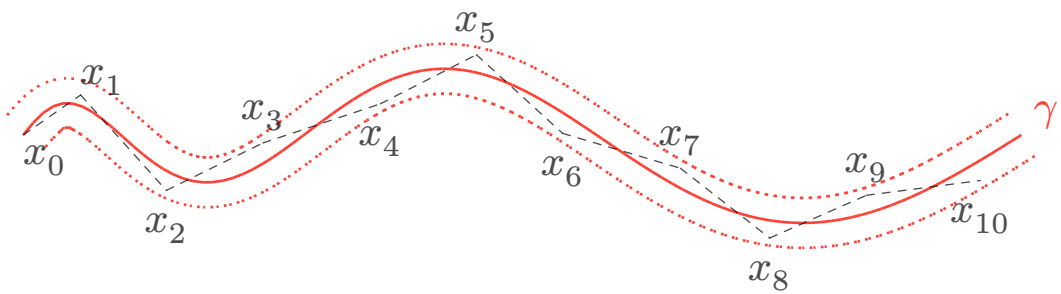
Cgf. of $Z(x; \omega)$: $\mathcal{H}(x, v) = \log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right]$

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

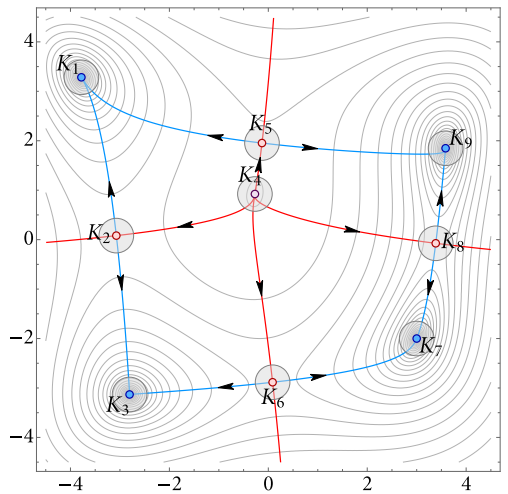
$$\dot{\gamma}_t = -\nabla f(\gamma_t)$$

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

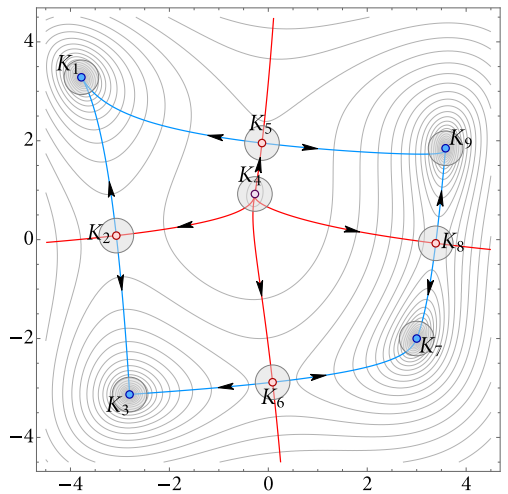
$$\dot{\gamma}_t = -\nabla f(\gamma_t)$$

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 father γ is from being a gradient flow, the smaller $\mathbb{P}(\text{SGD} \approx \gamma)$

Transition between critical points

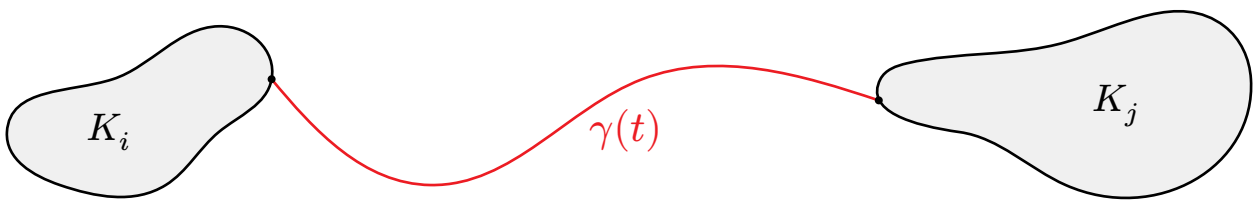
Given K_i, K_j components of critical points, what is $\mathbb{P}\big(\text{SGD transitions from } K_i \text{ to } K_j\big)$?

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:

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

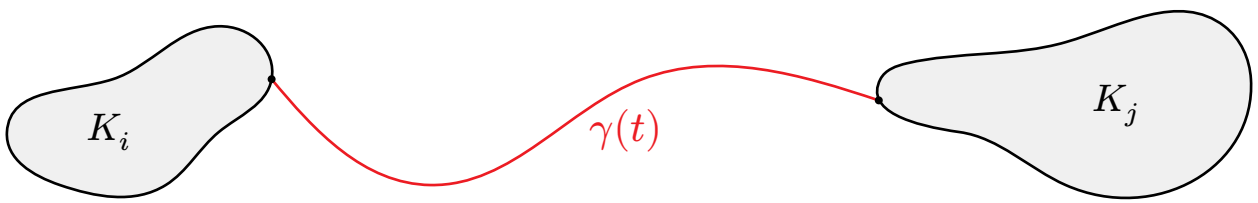


Transition between critical points

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:

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



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

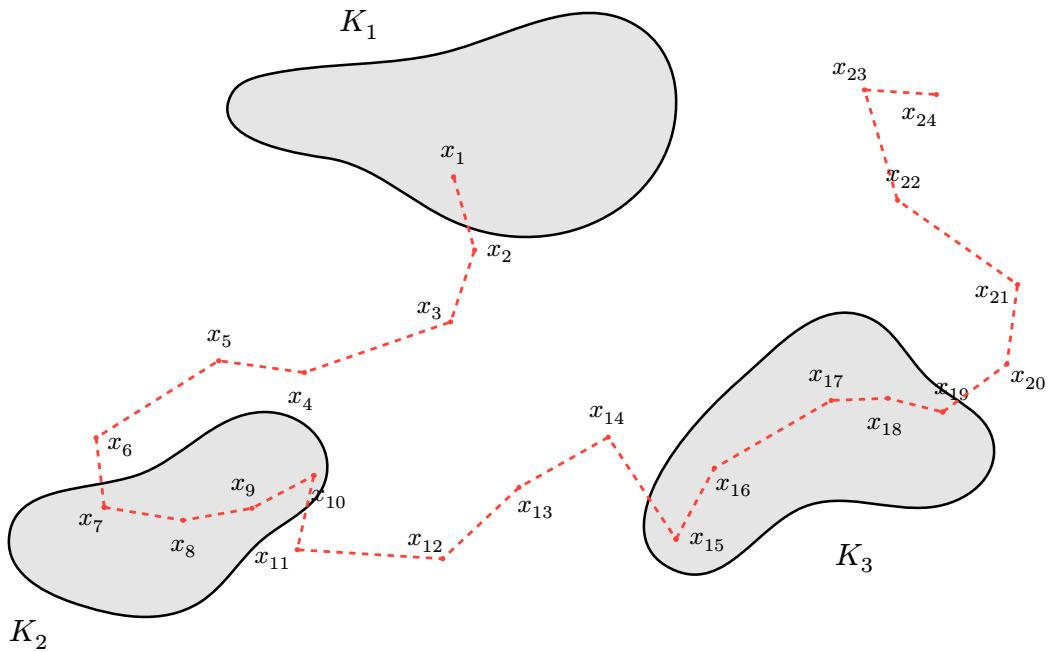
$$\mathbb{P}\big(\textit{SGD transitions from } K_i \textit{ to } K_j\big) \approx \exp\left(-\frac{B_{i,j}}{\eta}\right)$$

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:

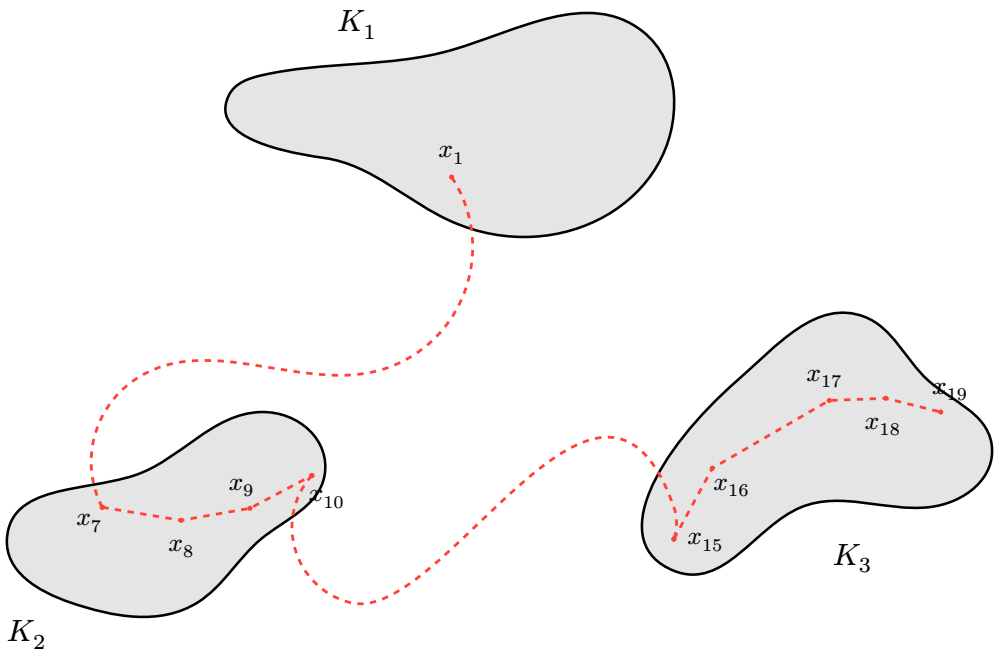
$$B_{i,j} \geq \frac{2\big(f(K_j) - f(K_i)\big)}{\sigma^2}$$

Restriction to critical components



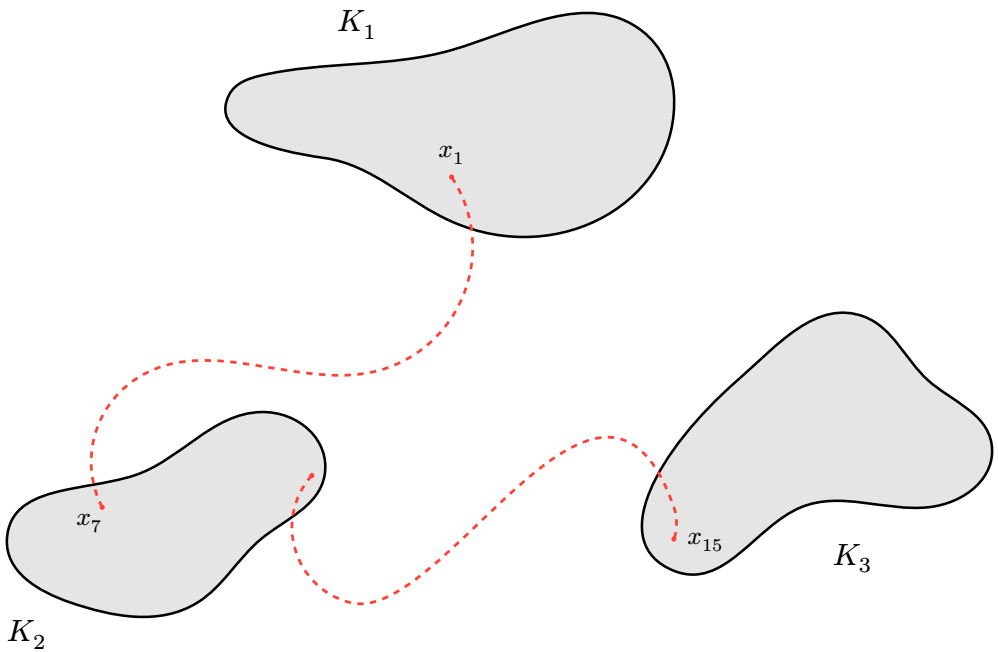
Main idea of the proof: Restrict SGD to a chain visiting only critical components

Restriction to critical components



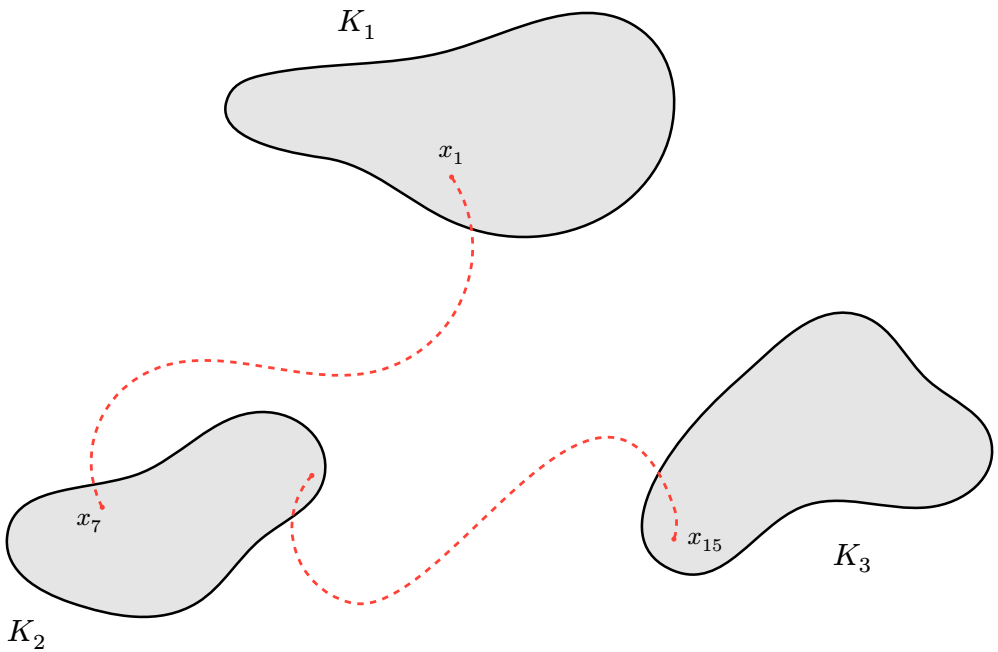
Main idea of the proof: Restrict SGD to a chain visiting only critical components

Restriction to critical components



Main idea of the proof: Restrict SGD to a chain visiting only critical components

Restriction to critical components



Main idea of the proof: Restrict SGD to a chain visiting only critical components

$\rightarrow$ study SGD as a finite-state space Markov chain on $\{K_1, ..., K_p\}$ with

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

Energy

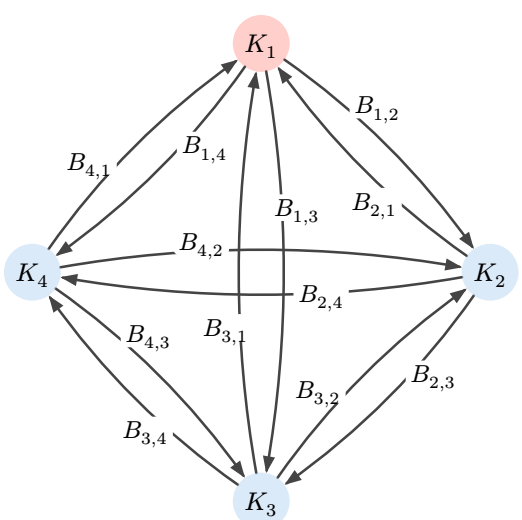
Using exact formulas for finite-state space Markov chains:

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*:*

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



Energy

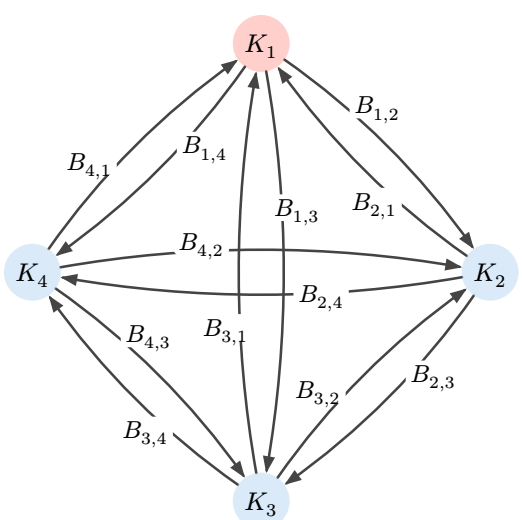
Using exact formulas for finite-state space Markov chains:

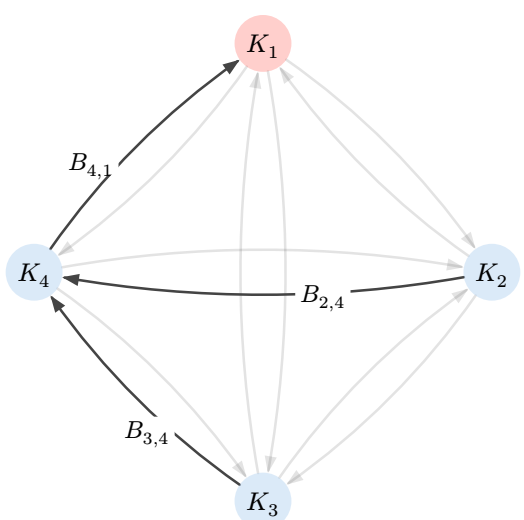
Lemma *(very informal)**:** the invariant measure of SGD restricted to* $\{K_1, ..., K_p\}$ *is, for* $\eta > 0$ *small enough,*

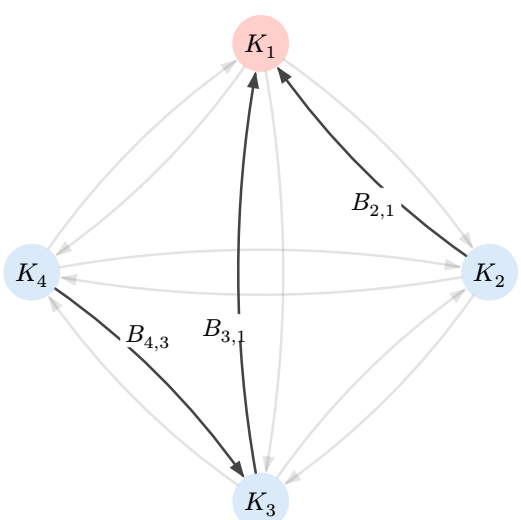
$$\pi(i) \propto \exp\left(-\frac{E_i}{\eta}\right)$$

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

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







Main Theorem

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:*** *there is some $c > 0$ s.t.*

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

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 $K_0 = \operatorname{argmin}_i E_i$*

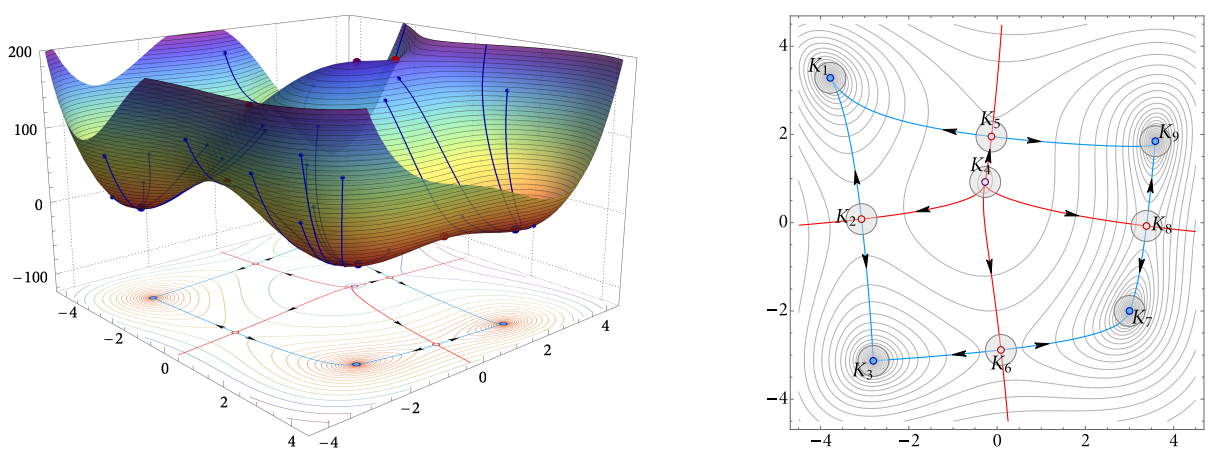
$$\mu_\infty(\mathcal{U}_0) \geq 1 - e^{-\frac{c}{\eta}}, \qquad \text{for some } c > 0$$

Example: Gaussian noise

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

Boltzmann-Gibbs distribution: for all i,

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



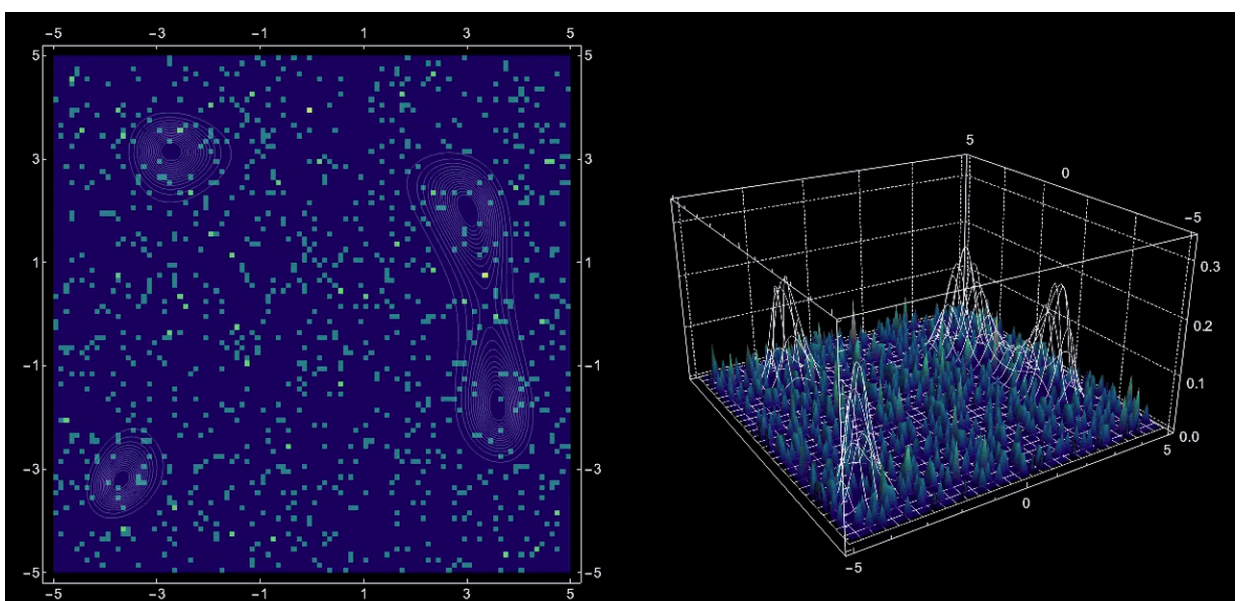
Himmelblau function

Example: Gaussian noise

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

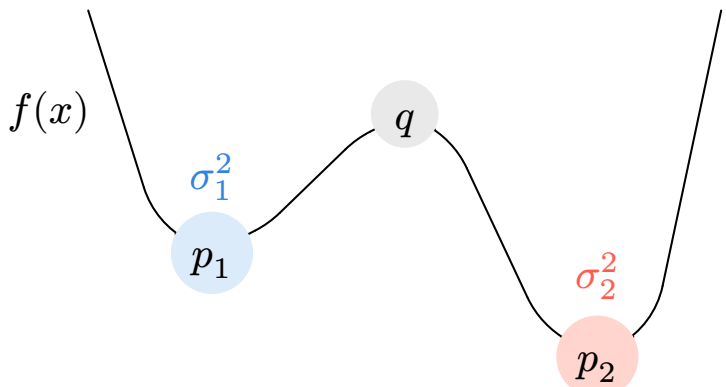
Boltzmann-Gibbs distribution: for all i,

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

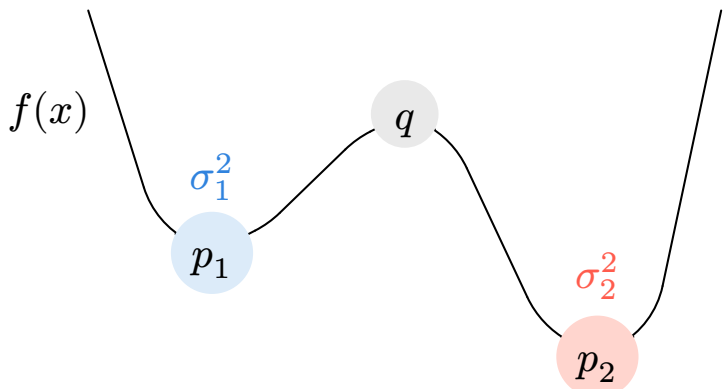


Evolution of the distribution of the iterates of SGD, initialized at random

Minimizers of the energy = minimizers of the function?

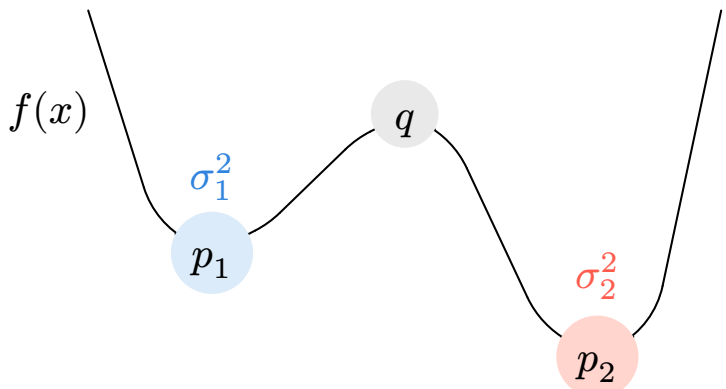


Minimizers of the energy = minimizers of the function?



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

Minimizers of the energy = minimizers of the function?



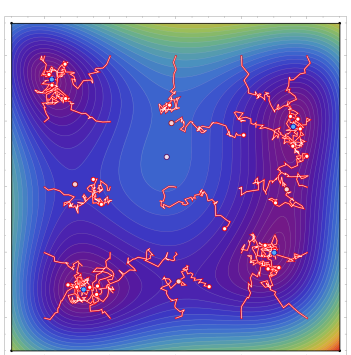
$$E_1 = \frac{f(q) - f(p_2)}{\sigma_2^2} \qquad \text{and} \qquad E_2 = \frac{f(q) - f(p_1)}{\sigma_1^2}$$

If σ_1 small enough, $E_1 < E_2$ and so $\mu_\infty(p_1) \ll \mu_\infty(p_2)$ even if x_1 is not a global minimizer!

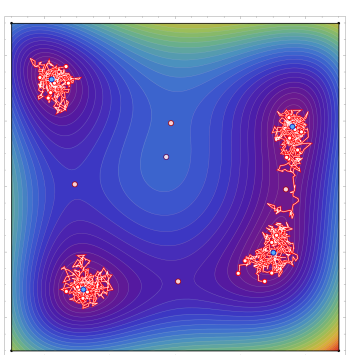
$\rightarrow$ In general, minimizer of the energy $\neq$ minimizer of the function!

Conclusion

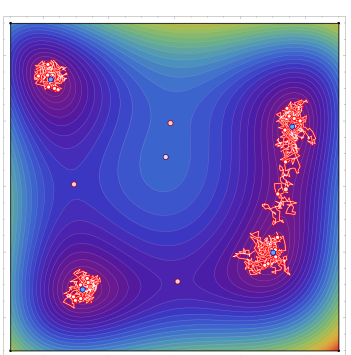
- We characterized the asymptotic distribution of SGD: it concentrates near local minima, and in particular near those that minimize the energy.
- For this, we developed a new theoretical framework to analyze the long-run behavior of SGD in non-convex landscapes through large deviations.



$t = 50$



$t = 200$

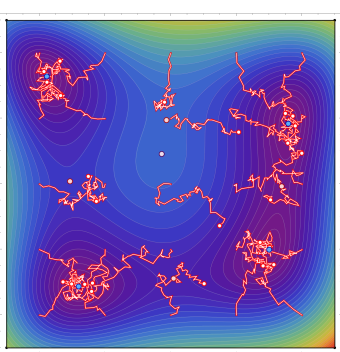


$t = 500$

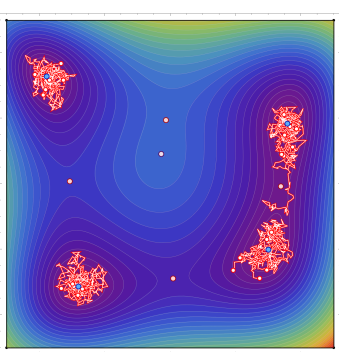
arXiv:2406.09241

Conclusion

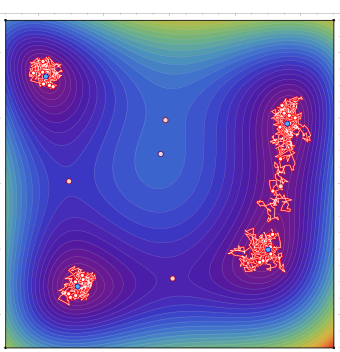
- We characterized the asymptotic distribution of SGD: it concentrates near local minima, and in particular near those that minimize the energy.
- For this, we developed a new theoretical framework to analyze the long-run behavior of SGD in non-convex landscapes through large deviations.
- More broadly, it opens the door to analyzing stochastic non-convex optimization algorithms beyond SGD, including Adam and min-max dynamics.
- We analyzed the asymptotic distribution but our framework allows much more: see Jérôme's talk!



$t = 50$



$t = 200$



$t = 500$

arXiv:2406.09241

What is known?

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

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

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 can't get no information about which one.

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

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

  Convergence of the distribution of the iterates to e^{-f} (Raginsky et al., 2017) but scaling of the noise differs from SGD
  $\Rightarrow$ analysis does not carry over

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

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

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