What is the Long-Run Behaviour of SGD?

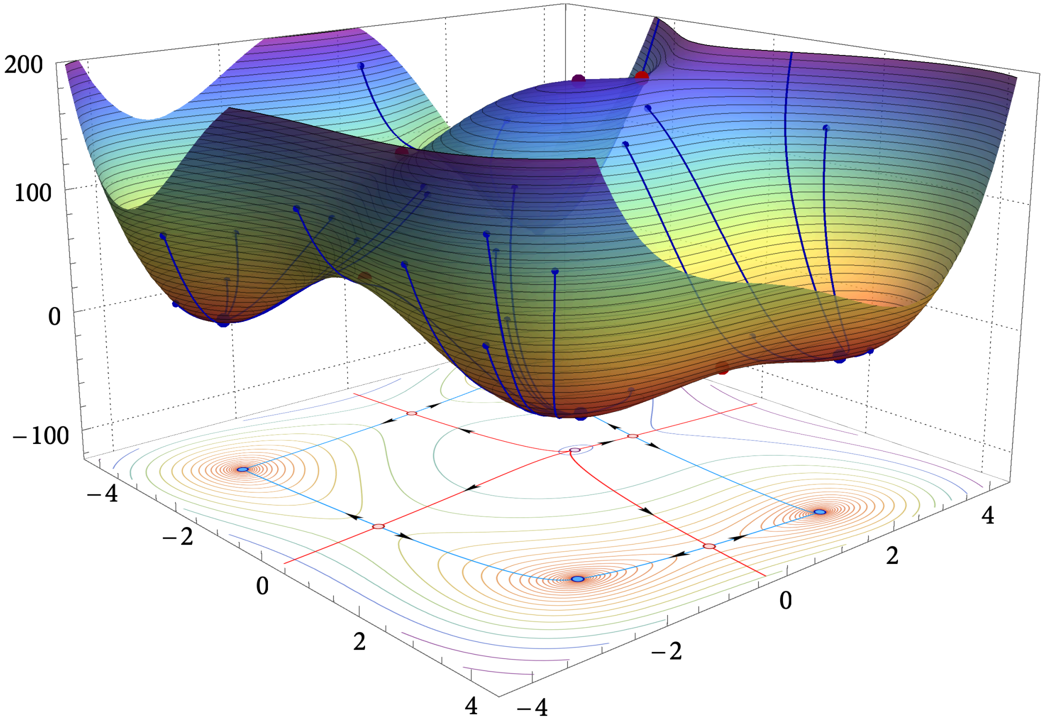
Morgan Stanley ML Research Talk

December 10, 2025

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

Himmelblau function

$$f(x,y) = \left(x^2 + y - 11\right)^2 + \left(x + y^2 - 7\right)^2$$



Himmelblau function

Training in machine learning = stochastic gradient methods

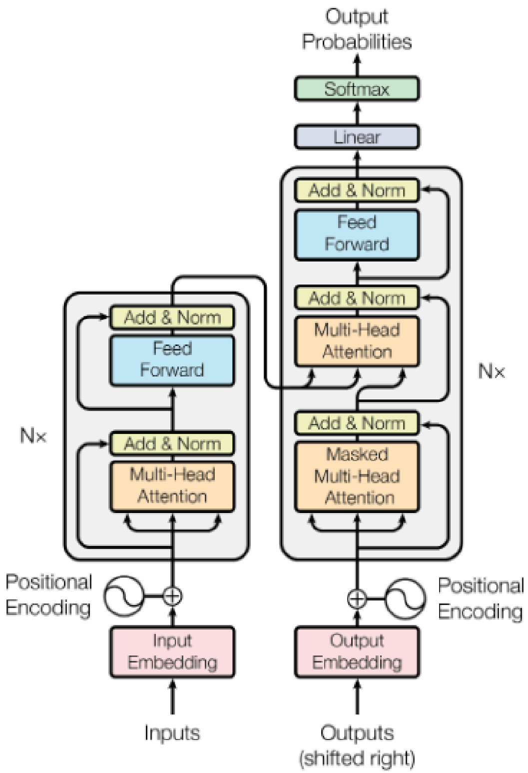
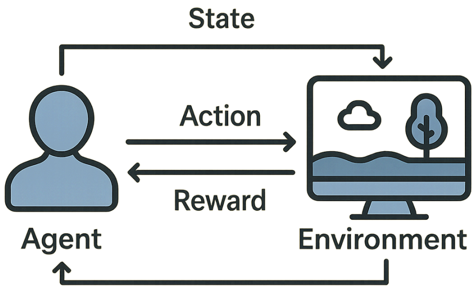
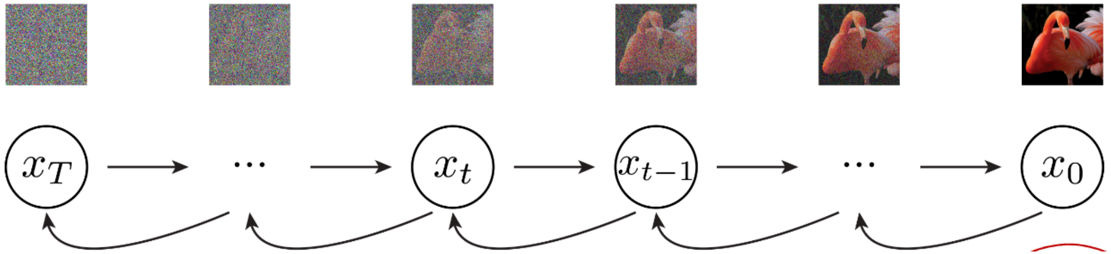


Image credit: Vaswani et al., 2017

Image credit: Meta AI





Different domains, same training method = stochastic gradient methods

Nonconvex loss!

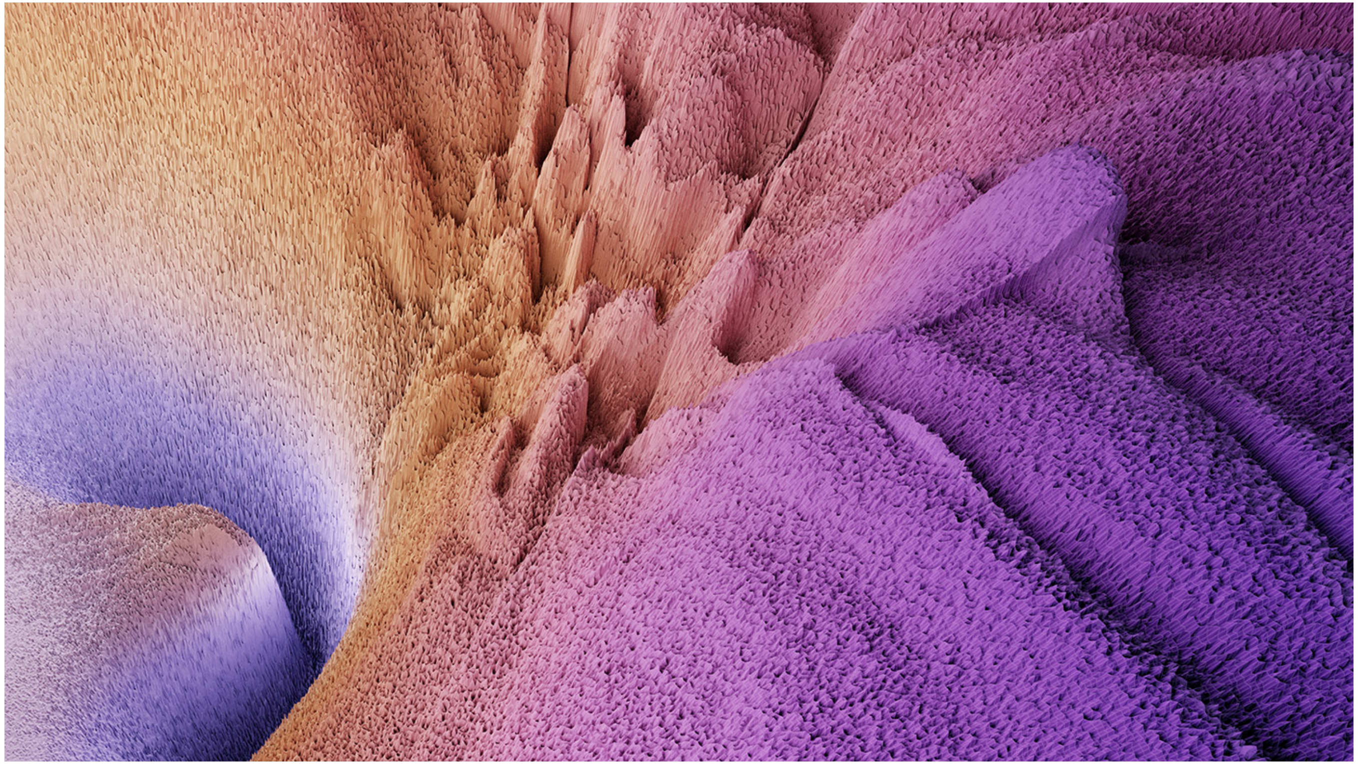


Image credit: losslandscape.com

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

Core focus: SGD

For $f : \mathbb{R}^d \to \mathbb{R}$ loss of model with parameters x,

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\, f(x) \quad \text{where} \qquad f(x) = \mathbb{E}_\omega[f(x;\omega)]$$

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

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

Core focus: SGD

For $f : \mathbb{R}^d \to \mathbb{R}$ loss of model with parameters x,

$$\operatorname*{minimize}_{x \in \mathbb{R}^d} f(x) \quad \text{where} \qquad f(x) = \mathbb{E}_\omega[f(x;\omega)]$$

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:

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

$$\begin{aligned} x_{t+1} &= x_t - \eta \nabla f_{i_t}(x_t) \\ &= x_t - \eta \Big[\nabla f(x_t) + \underbrace{\nabla f_{i_t}(x_t) - \nabla f(x_t)}_{\text{zero-mean noise}}\Big] \end{aligned}$$

Core focus: SGD

For $f : \mathbb{R}^d \to \mathbb{R}$ loss of model with parameters x,

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x) \quad \text{where} \quad f(x) = \mathbb{E}_\omega[f(x;\omega)]$$

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

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

step-size   zero-mean noise

Q: What is the asymptotic behavior of SGD?

Core focus: SGD

For $f : \mathbb{R}^d \to \mathbb{R}$ loss of model with parameters x,

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\; f(x) \quad \text{where} \quad f(x) = \mathbb{E}_\omega[f(x;\omega)]$$

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?

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

$\to$ **Q2:** How much time to get there?

What is known?

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 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

What is known?

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

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

SGD with constant step-size:

- f strongly convex: SGD converges near the minimizer (Polyak, 1987)
- f convex: average of SGD iterates (almost) optimal (Polyak & Juditsky, 1992)
- 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 (Brandière & Duflo, 1996; Mertikopoulos et al., 2020)

$\to$ **Q1:** Which critical points (and which local minima) are visited most often in the long run?

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

New approach: large deviations

TLDR: we describe the asymptotic behavior of SGD in nonconvex problems through a large deviation approach

Outline:
1. Introduction
2. Asymptotic distribution of SGD
3. Global convergence time of SGD

Based on our papers:
- *What is the long-run behavior of SGD? A large deviation analysis.* ICML 2024
- *The global convergence time of SGD in non-convex landscapes.* ICML 2025

On the objective function f

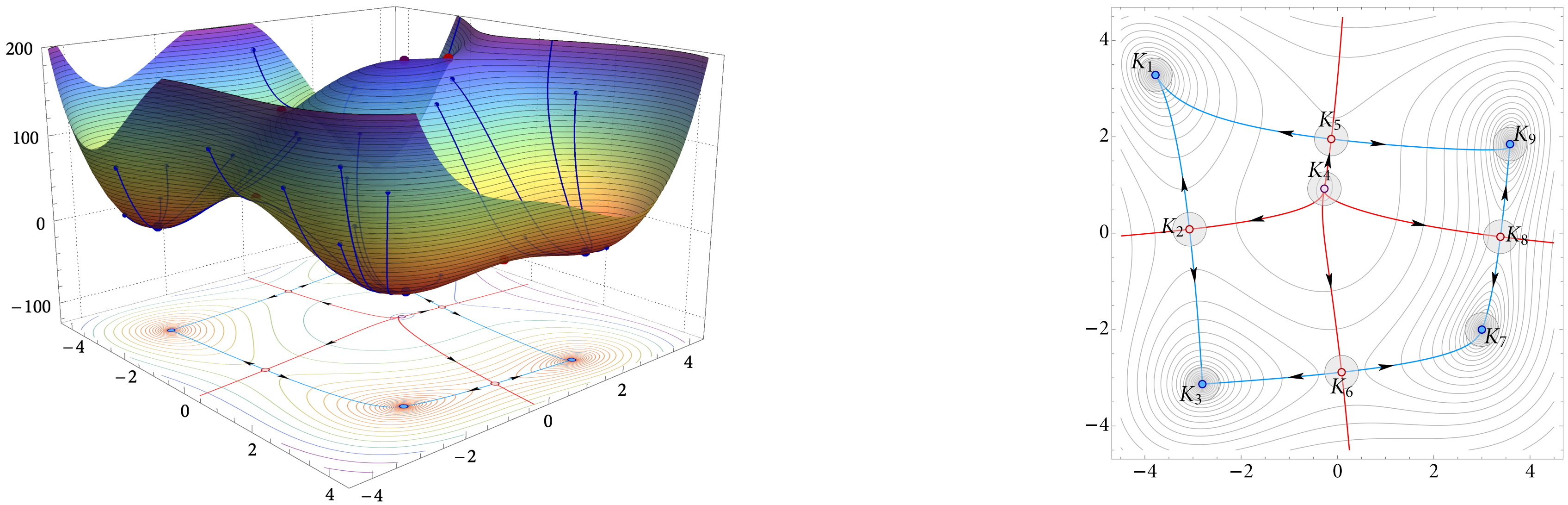
Regularity assumption:

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

where K_i connected components (compact)

$\rightarrow$ Avoids pathological cases, realistic in practice

Himmelblau function



Asymptotic distribution of SGD

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

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

Invariant measures are limit points of the mean occupation measures of the iterates of SGD:

for any set $\mathcal{B}$ of interest, as $n \to \infty$,

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

Invariant measure: probability measure μ_∞ such that

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

Q1: Where do invariant measures of SGD concentrate?

Main results (informal)

Recall:

$$\mathrm{crit}(f) := \{x : \nabla f(x) = 0\} = \{K_1, K_2, ..., K_p\} \text{ with } K_i \text{ connected components}$$

1. **Concentration near critical points:**

$$\mu_\infty(\mathrm{crit}(f)) \to 1 \quad \text{as } \eta \to 0$$

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,

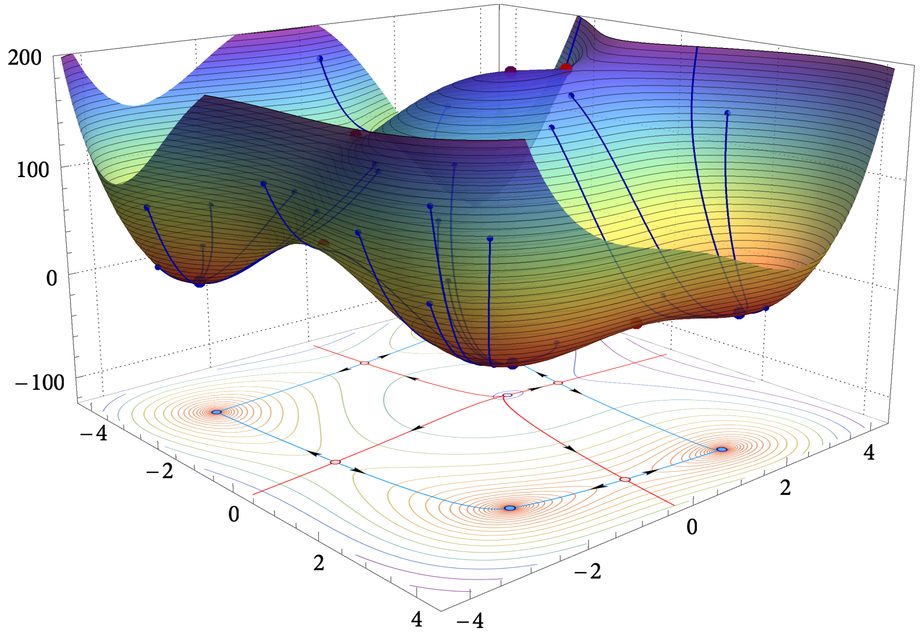
$$\mu_\infty(K_i) \propto \exp\left(-\frac{E_i}{\eta}\right)$$

4. **Ground state concentration:** there is K_{i_0} that minimizes E_i such that,

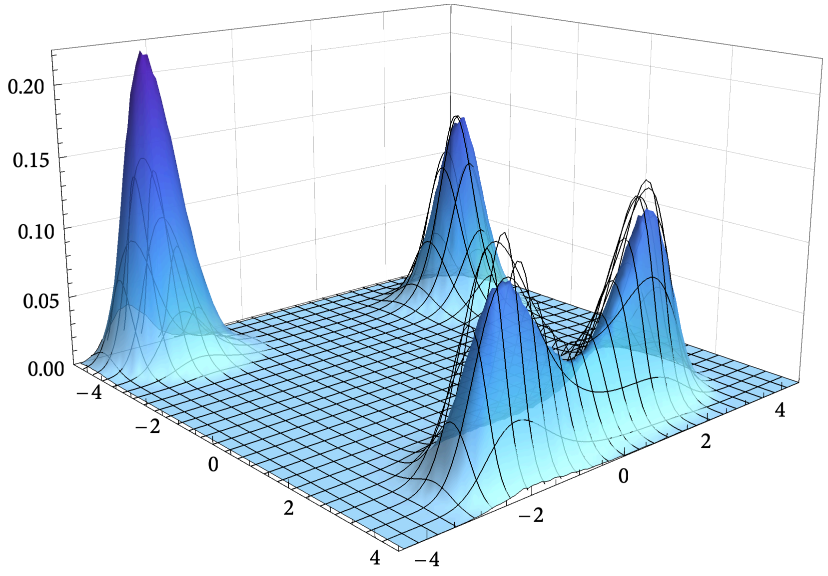
$$\mu_\infty\left(K_{i_0}\right) \to 1 \quad \text{as } \eta \to 0$$

Example: Gaussian noise

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



Himmelblau function



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

Global convergence time of SGD

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

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

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

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

Main result (informal)

Global convergence time of SGD: starting at x, the time τ to reach $\operatorname{argmin} f$ satisfies

$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{J(x)}{\eta}\right)$$

where $J(x)$ energy of SGD starting at x, for η, δ small enough

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

- The difficulty of the loss landscape
- The statistics of the noise

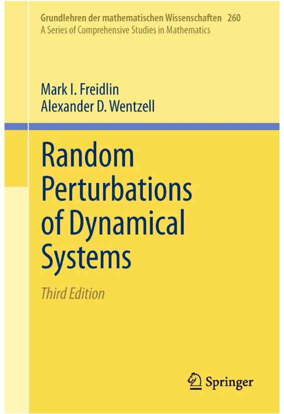
Challenges and techniques

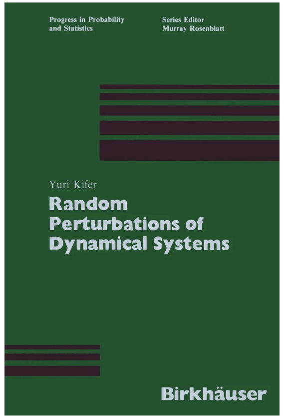
- No known approach to analyze the asymptotic distribution of SGD on non-convex problems
- We leverage large deviation theory and the theory of random perturbations of dynamical systems,
  $\rightarrow$ Estimate the probability of rare events, such as SGD escaping a local minima
- We adapt the theory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

References

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

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





Objective and noise assumptions

Recall: we assume

$$\operatorname{crit}(f) := \{x : \nabla f(x) = 0\} = \{K_1, K_2, ..., K_p\}$$

where K_i connected components (compact)

Objective assumptions:

- ∇f is Lipschitz-continuous
- f is coercive: $\lim_{\|x\| \to \infty} f(x) = \lim_{\|x\| \to \infty} \|\nabla f(x)\| = +\infty$

Noise assumptions:

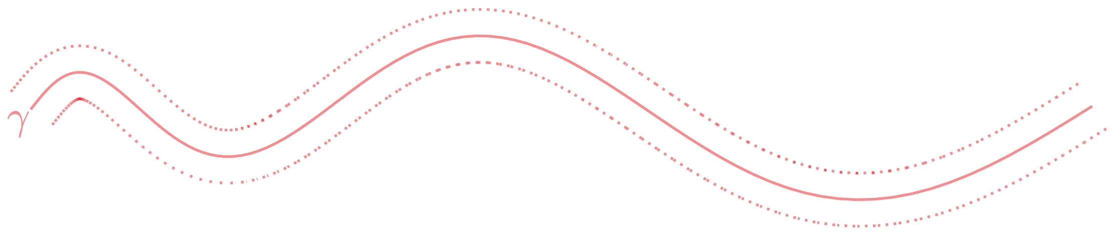
- $\mathbb{E}[Z(x;\omega)] = 0$, $\operatorname{cov}(Z(x;\omega)) \succ 0$, $Z(x;\omega) = O(\|x\|)$ almost surely
- $Z(x;\omega)$ is σ sub-Gaussian:

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

$\rightarrow$ Realistic in the context of deep learning (normalization layers, weight decay, GeLU/Swish activations, etc.)

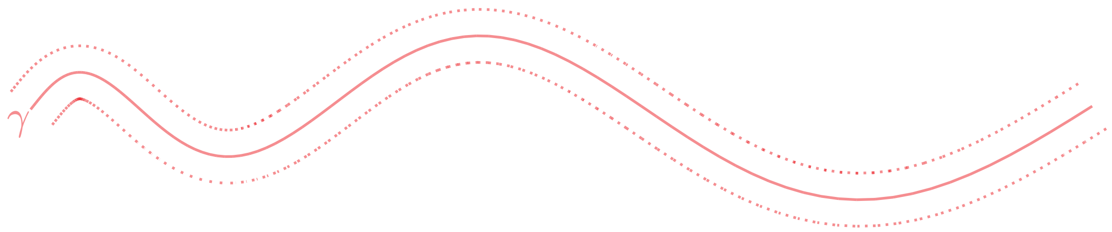
Large deviations for discrete-time SGD

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 $\gamma : [0, T] \to \mathbb{R}^d$,*

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

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

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

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

LDP in the Gaussian case

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

Cumulant generating function: $\mathcal{H}(x, v) = \frac{\sigma^2}{2}\|v\|^2$

Lagrangian: $\mathcal{L}(x, v) = \frac{\|v+\nabla f(x)\|^2}{2\sigma^2}$

Action functional:

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

LDP in the Gaussian case

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

Cumulant generating function: $\mathcal{H}(x, v) = \frac{\sigma^2}{2}\|v\|^2$

Lagrangian: $\mathcal{L}(x, v) = \frac{\|v+\nabla f(x)\|^2}{2\sigma^2}$

Action functional:

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

Key observations:

- γ is a trajectory of a gradient flow trajectory: $\dot{\gamma}_t = -\nabla f(\gamma_t)$ iff $\mathcal{S}_T[\gamma] = 0$
- The farther γ is from being a gradient flow, the larger $\mathcal{S}_T[\gamma]$
- And, as a consequence, the smaller the probability of SGD following γ:

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

Transition between critical points

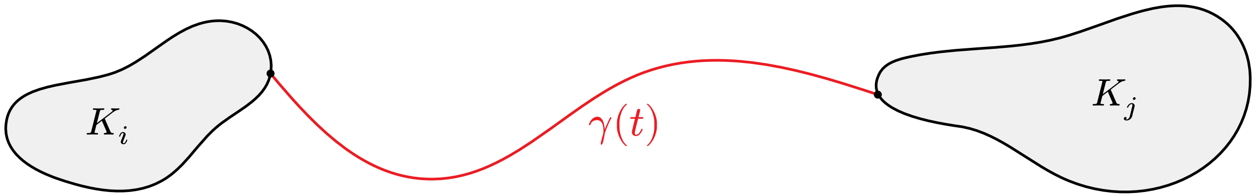
Given K_i, K_j critical points, what is $\mathbb{P}(\text{SGD transitions from } K_i \text{ to } K_j)$?

Transition between critical points

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

Involves the transition cost:

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



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:

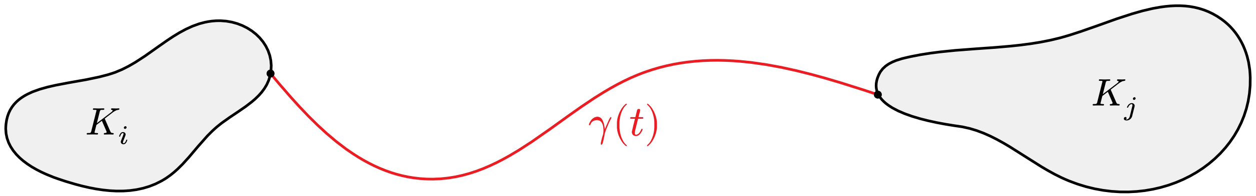
$$B_{i,j} \geq \frac{2(f(K_j) - f(K_i))}{\sigma^2}$$

Transition between critical points

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

Involves the transition cost:

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



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:

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

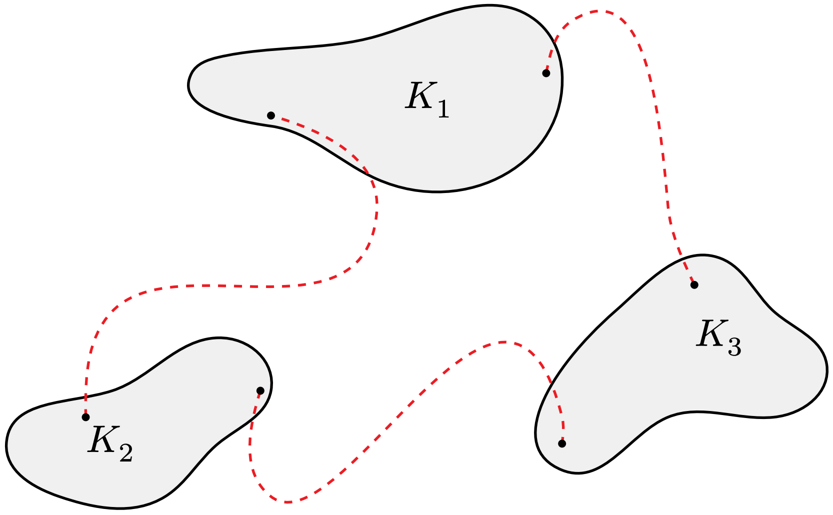
Proposition: *Transition probability from K_i to K_j: for $\eta > 0$ small enough,*

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

Restriction to critical components

Recall:

$$\text{crit}(f) := \{x : \nabla f(x) = 0\} = \{K_1, K_2, ..., K_p\} \text{ with } K_i \text{ connected components}$$



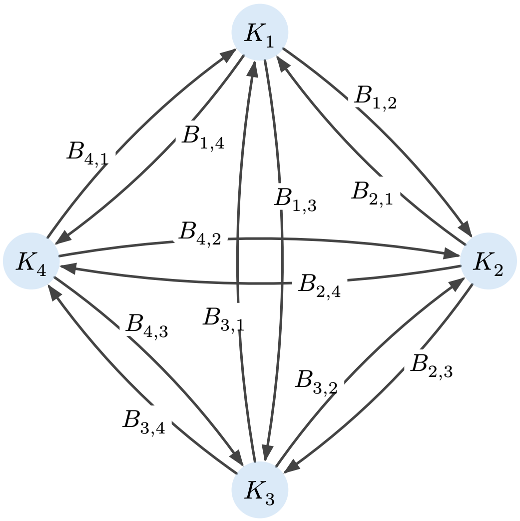
Main idea of the proof: Restrict SGD to a chain visiting only critical components

$\rightarrow$ studies a chain on $\{1, ..., p\}$

Transition graph

Study SGD as a Markov chain on $\{1, ..., p\}$ with transitions

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



Transition graph: complete graph on $\{1, ..., p\}$ with weights $B_{i,j}$ on $i \to j$

$\to$ leverage exact formulas for finite-state space Markov chains

Energy

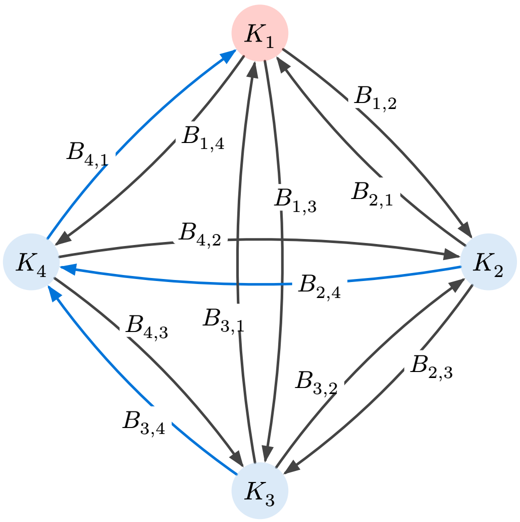
Using exact formulas for finite-state space Markov chains:

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

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

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 results (more formal)

Recall:

$$\mathrm{crit}(f) := \{x : \nabla f(x) = 0\} = \{K_1, K_2, ..., K_p\} \text{ with } K_i \text{ connected components}$$

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

1. ***Concentration on*** $\mathrm{crit}(f)$***:*** *there is some* $\lambda > 0$ *s.t.*

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

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

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

3. ***Avoidance of non-minimizers:*** *if* K_i *is not minimizing, there is* K_j *minimizing with* $E_j < E_i$*:*

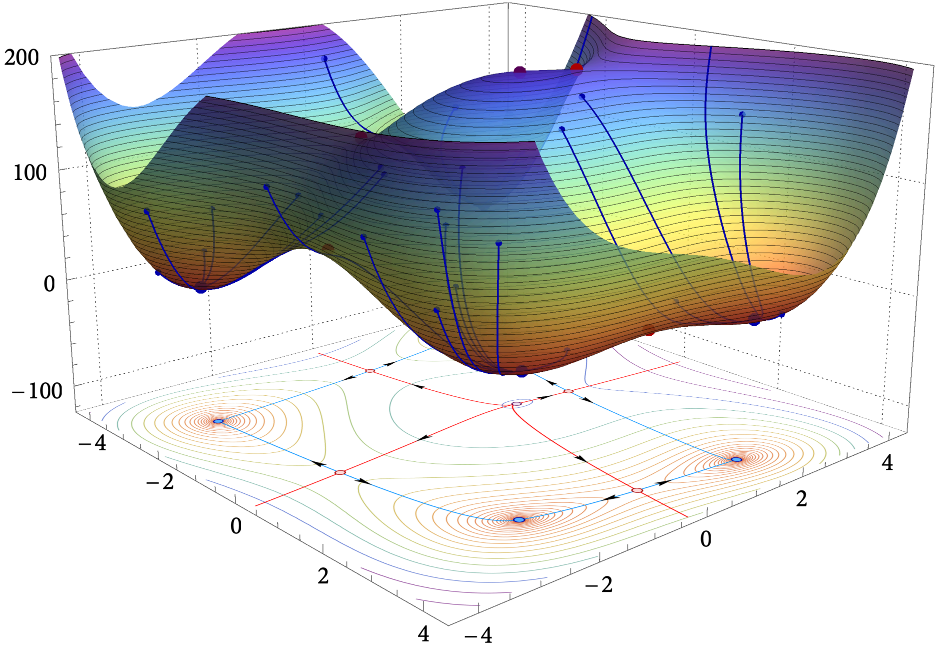
$$\frac{\mu_\infty(\mathcal{U}_i)}{\mu_\infty(\mathcal{U}_j)} \leq e^{-\frac{\lambda_{i,j}}{\eta}} \qquad \text{for some } \lambda_{i,j} > 0$$

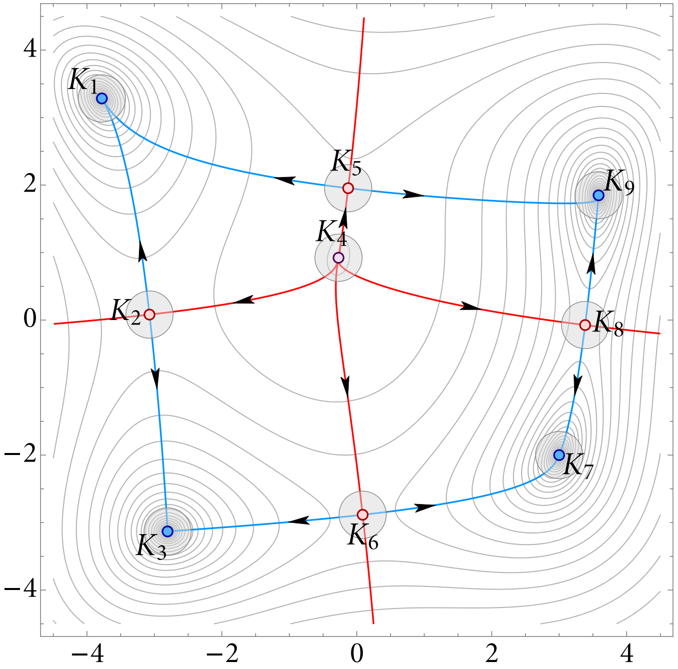
4. ***Concentration on ground states:*** *given* $\mathcal{U}_0$ *neighborhood of the ground states* $K_0 = \mathrm{argmin}_i\, E_i$

$$\mu_\infty(\mathcal{U}_0) \geq 1 - e^{-\frac{\lambda_0}{\eta}}, \qquad \text{for some } \lambda_0 > 0$$

Example: Gaussian noise

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

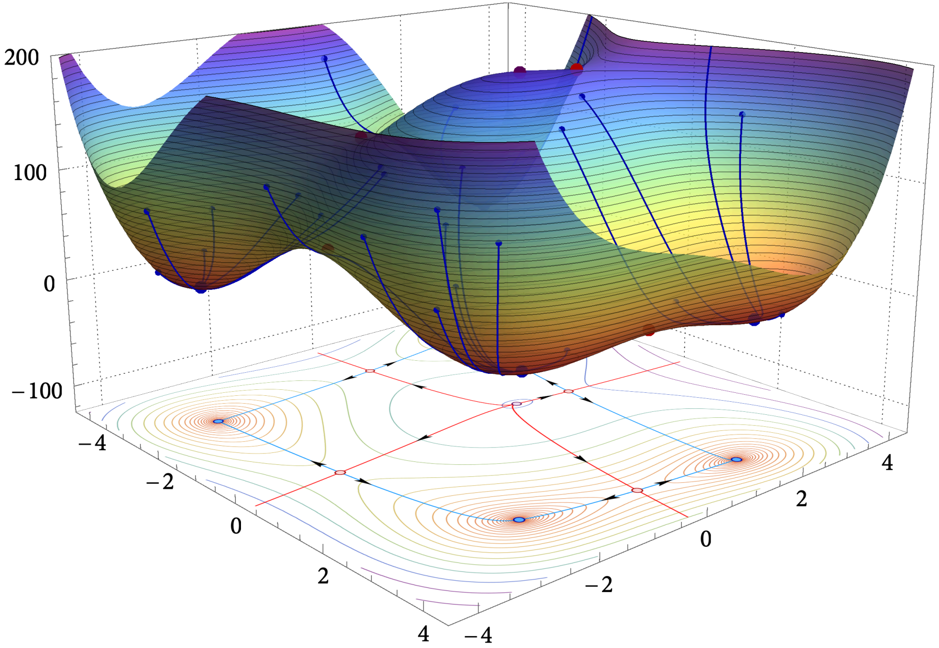


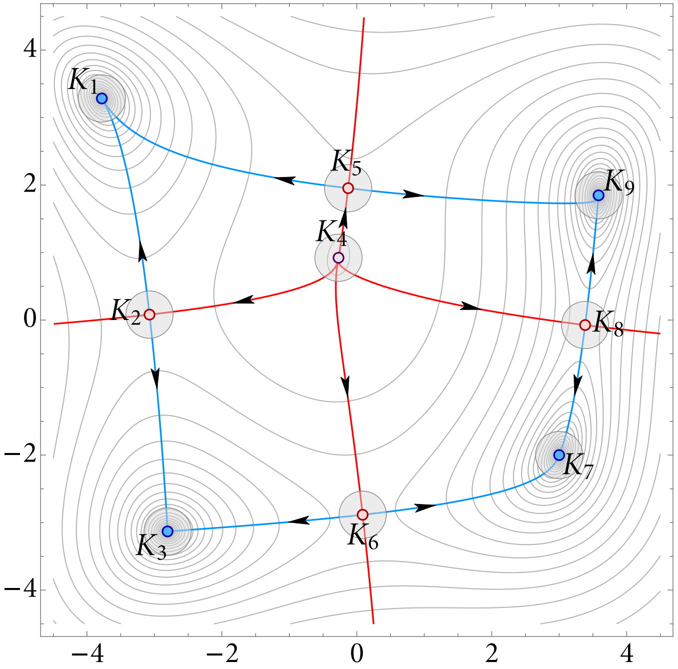


$$B_{5,1} = 0; \qquad B_{1,5} = \frac{2(f(K_5) - f(K_1))}{\sigma^2}$$

Example: Gaussian noise

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

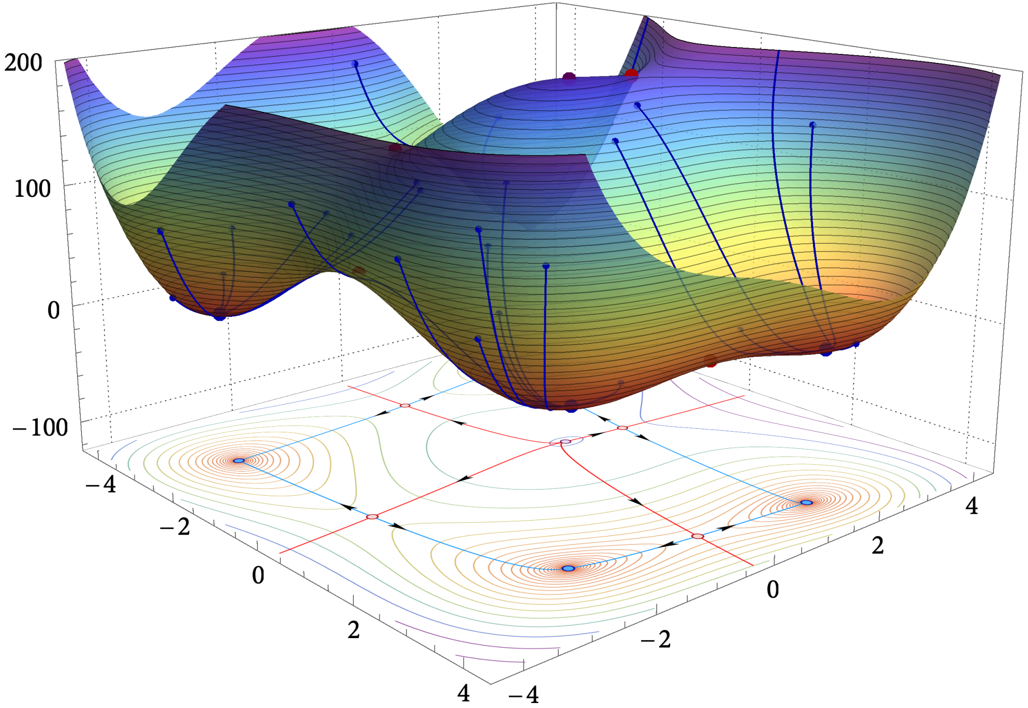




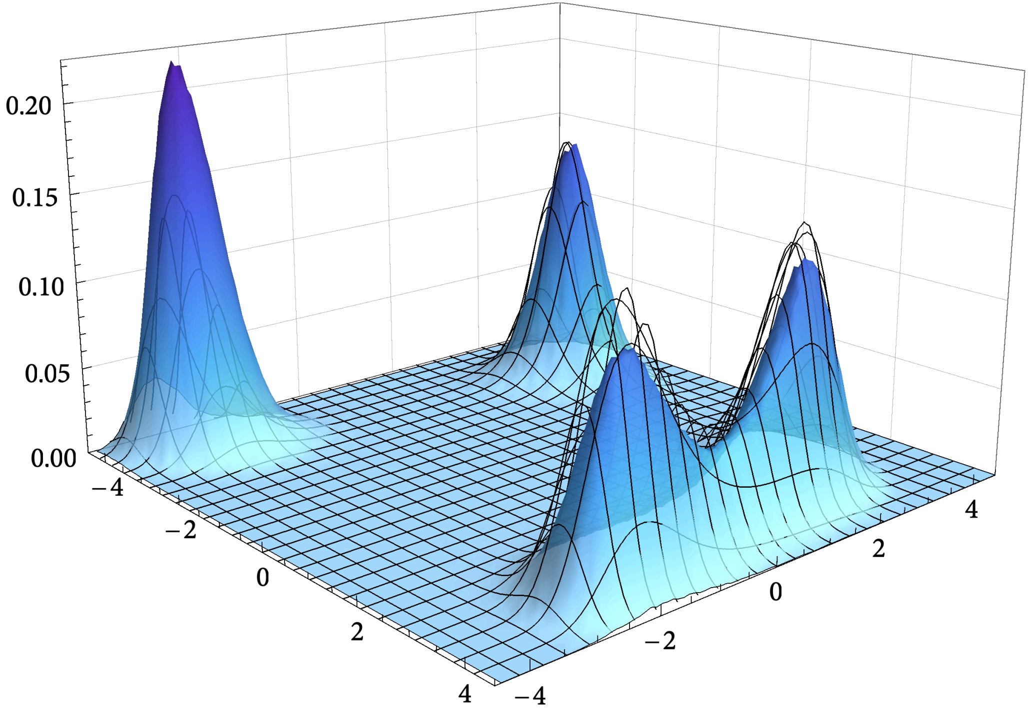
$$E_i = \frac{2f(K_i)}{\sigma^2} \qquad \text{and} \qquad \mu_\infty(K_i) \approx \exp\left(-\frac{2f(K_i)}{\sigma^2 \eta}\right)$$

Example: Gaussian noise

If $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I_d)$, then $E_i = \frac{2f(K_i)}{\sigma^2}$

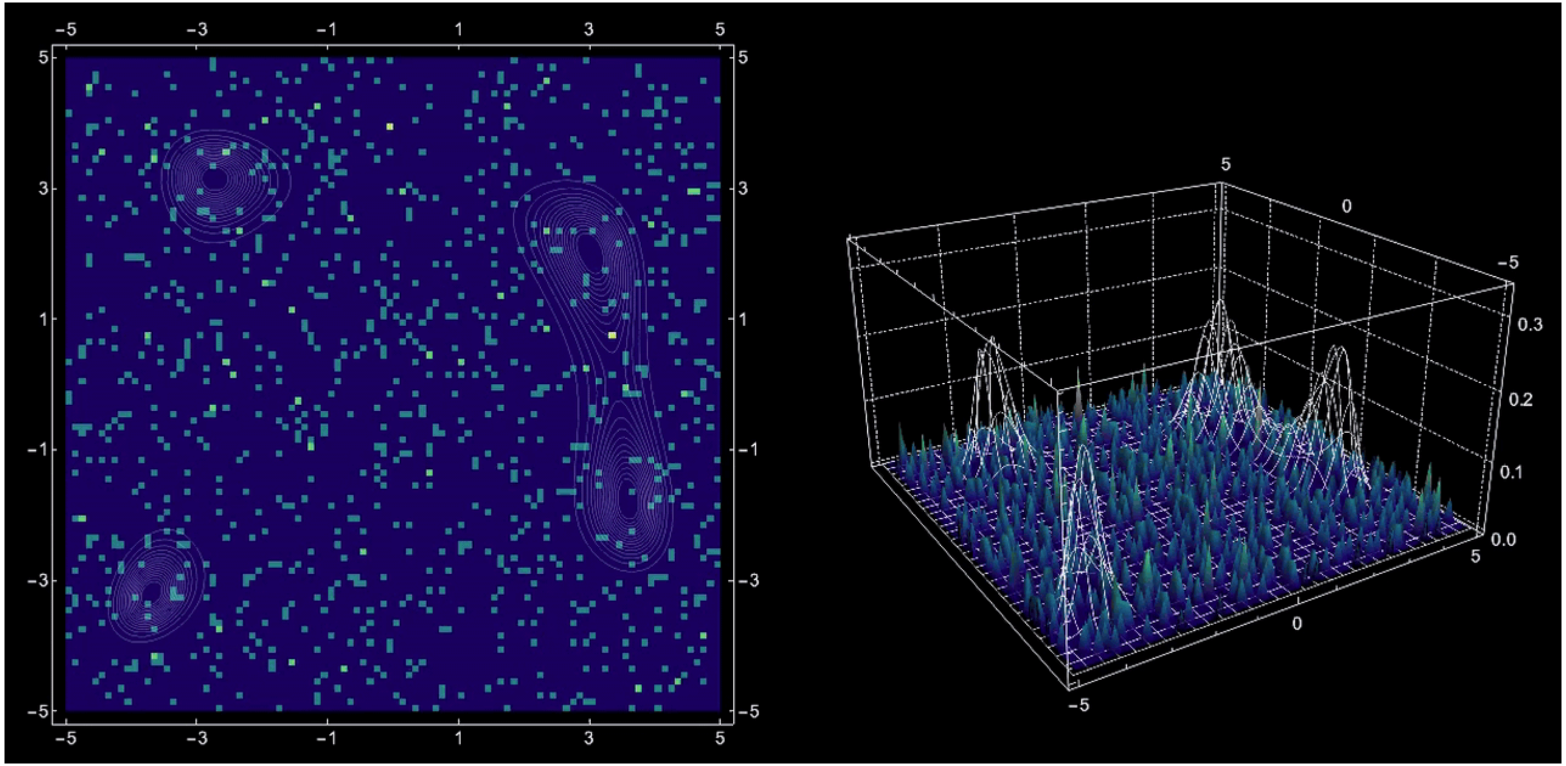


Himmelblau function



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

Example: Gaussian noise



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

Gaussian noise: general case

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

Boltzmann-Gibbs distribution: for all i,

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

Gaussian noise: general case

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

Boltzmann-Gibbs distribution: for all i,

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

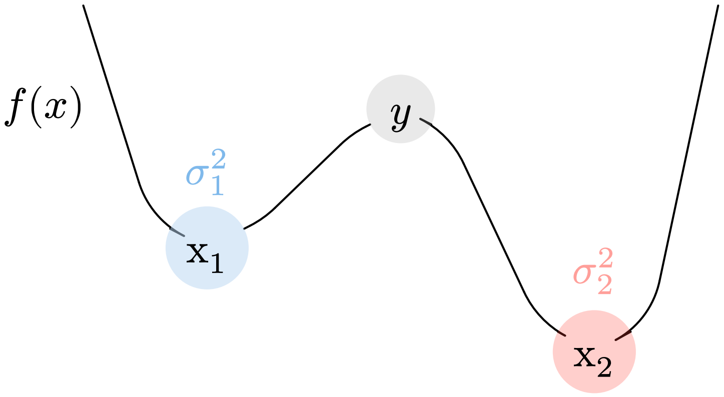
- Assuming $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 f(x) I_d)$

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

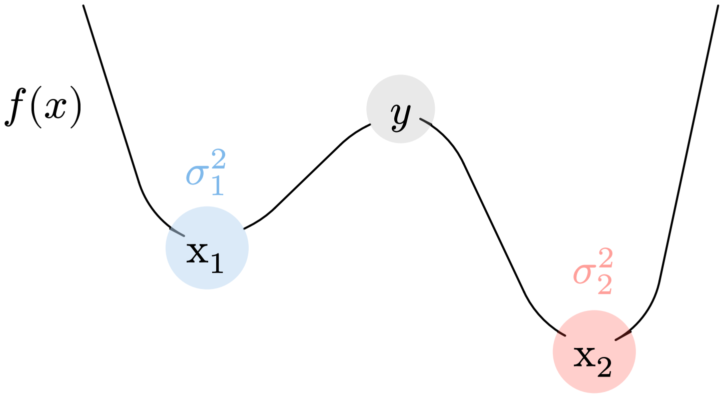
Power-law Gibbs distribution: for all i,

$$E_i = \frac{2\log f(K_i)}{\sigma^2} \qquad \text{and} \qquad \mu_\infty(K_i) \approx f(K_i)^{-\frac{2}{\sigma^2\eta}}$$

Minimizers of the energy = minimizers of the function?

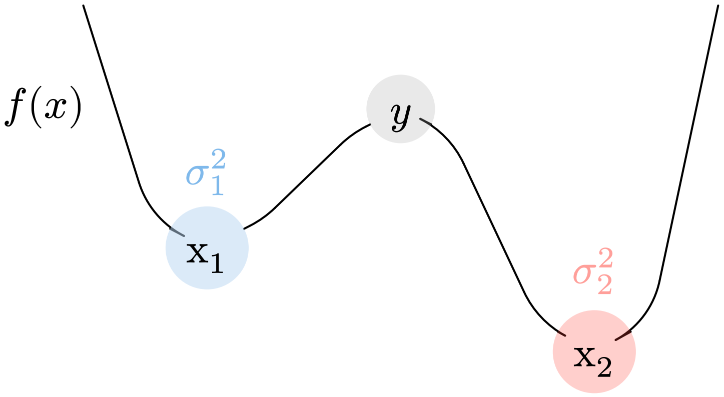


Minimizers of the energy = minimizers of the function?



$$E_1 = \frac{f(y) - f(x_2)}{\sigma_2^2} \quad \text{and} \quad E_2 = \frac{f(y) - f(x_1)}{\sigma_1^2}$$

Minimizers of the energy = minimizers of the function?



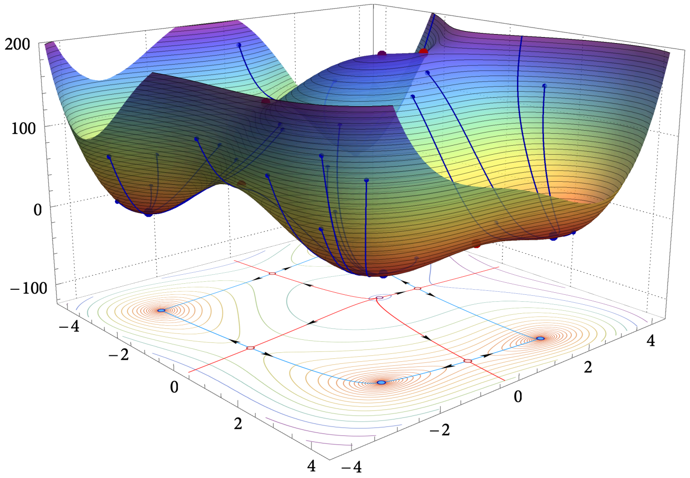
$$E_1 = \frac{f(y) - f(x_2)}{\sigma_2^2} \quad \text{and} \quad E_2 = \frac{f(y) - f(x_1)}{\sigma_1^2}$$

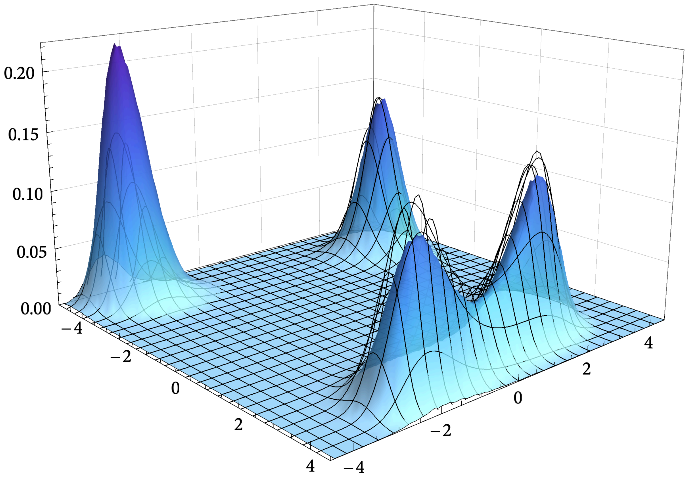
If σ_1^1 small enough, $E_1 < E_2$ and so $\mu_\infty(x_1) \ll \mu_\infty(x_2)$ even if x_1 is not a global minimizer!

$\rightarrow$ Question of the concentration of SGD remains intricate

Partial Conclusion (first part)

- We obtained a characterization of the invariant measure of SGD
- The relative weights of critical components depends on both the loss landscape and the noise structure
- Built on our large deviation framework to analyze the long-term behavior of SGD





Recall: global convergence time of SGD

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

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

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

Recall: global convergence time of SGD

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

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

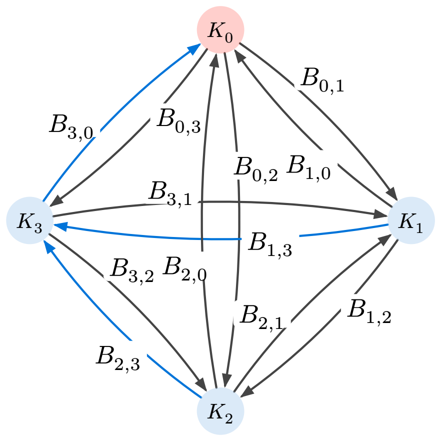
$$\tau = \min\{t \in \mathbb{N} \mid \operatorname{dist}(x_t, \operatorname{argmin} f) \leq \delta\}$$

Global convergence time of SGD: starting at x, the time τ to reach $\operatorname{argmin} f$ satisfies

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

where $J(x)$ “energy” of SGD starting at x, for any $\varepsilon > 0$ and η, $\delta > 0$ small enough

Definition of $J(x)$



Transition graph: *complete graph on* $\{0, \ldots, N-1\}$ *with weights* $B_{i,j}$ *on* $i \to j$

Energy of $K_0 = \operatorname{argmin} f$*:*

$$E_0 = \min\left\{ \sum_{j \to k \in T} B_{j,k} \mid T \text{ spanning tree pointing to } 0 \right\}$$

Energy of pruning K_i*:*

$$J(i \nrightarrow 0) = \min\left\{ \sum_{j \to k \in T} B_{j,k} \mid T \text{ spanning tree pointing to } 0 \text{ with an edge from } i \text{ to } 0 \text{ removed} \right\}$$

Energy of K_0 relative to K_i:

$$J(i) = E_0 - J(i \nrightarrow 0)$$

Energy of K_0 relative to x:

$$J(x) = \max_{i=1,\ldots,N-1} [J(i) - B(x,i)]_+$$

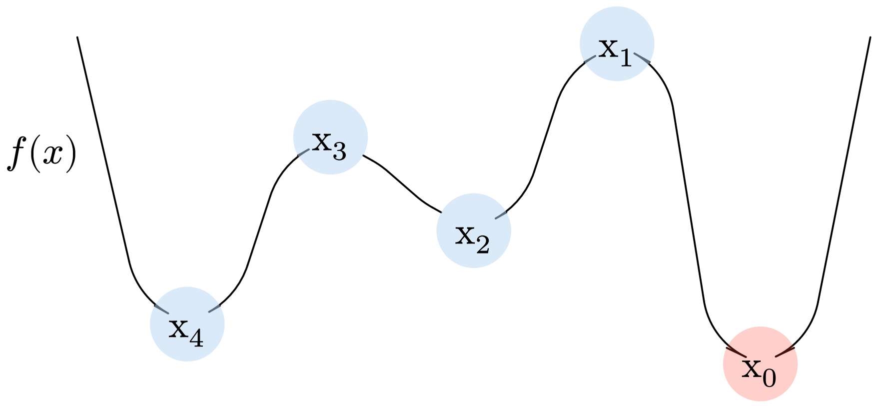
where $B(x,i)$ cost of the transition from x to K_i

$J(x)$: measure of the hardness of the problem

$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{J(x)}{\eta}\right)$$

General fact: $J(x) = 0$ for all x $\iff$ all local minima of f are global

$J(x) > 0$

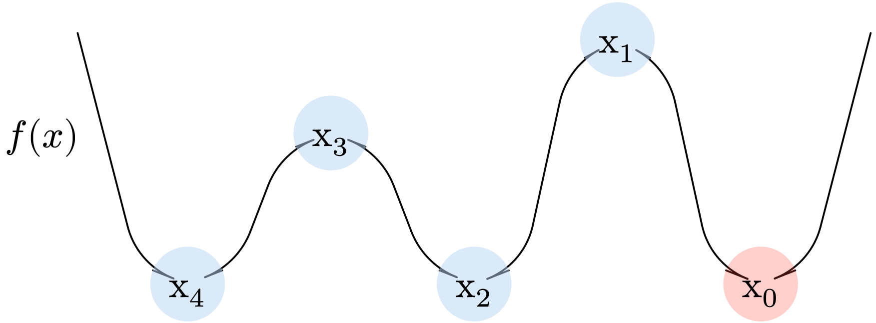


$J(x)$: measure of the hardness of the problem

$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{J(x)}{\eta}\right)$$

General fact: $J(x) = 0$ for all x $\iff$ all local minima of f are global

$J(x) = 0$



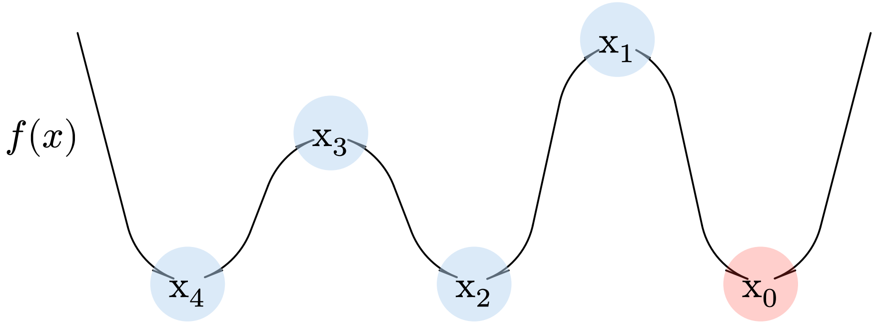
$J(x)$: measure of the hardness of the problem

$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{J(x)}{\eta}\right)$$

General fact: $J(x) = 0$ for all x $\iff$ all local minima of f are global

$\rightarrow$ neural networks when width $\geq$ # data points $+ 1$ (e.g. Nguyen et al., 2018; Nguyen et al., 2019)

$J(x) = 0$



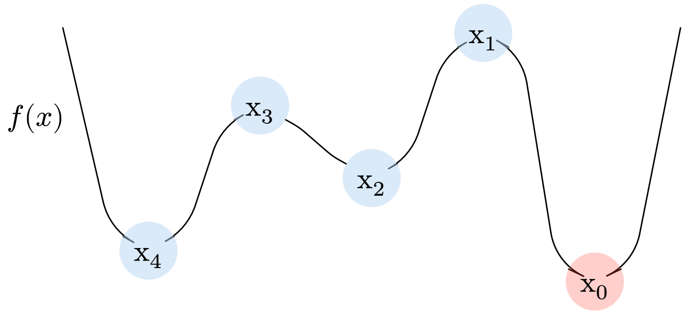
Gaussian bounds

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

Gaussian bound:

$$J(x) \leq \frac{2 \times \sharp\{\text{bad local minima}\} \times \{\text{max. saddle - min. bad local min.}\}}{\sigma^2}$$

$$J(x) \leq \frac{2 \times 2 \times (f(x_1) - f(x_4))}{\sigma^2}$$



Gaussian bounds

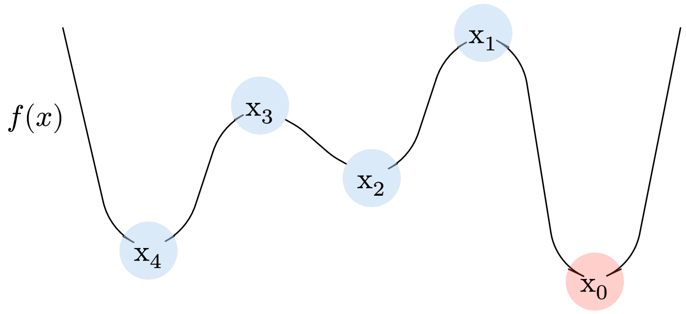
For Gaussian noise $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I_d)$,

Gaussian bound:

$$J(x) \leq \frac{2 \times \sharp\{\text{bad local minima}\} \times \{\text{max. saddle - min. bad local min.}\}}{\sigma^2}$$

$\rightarrow$ can be bounded as a function of width / depth of neural networks (e.g. Nguyen et al., 2021)

$$J(x) \leq \frac{2 \times 2 \times (f(x_1) - f(x_4))}{\sigma^2}$$



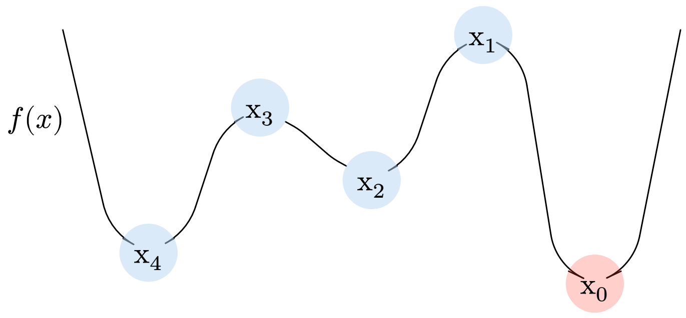
Power-law Gaussian bounds

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

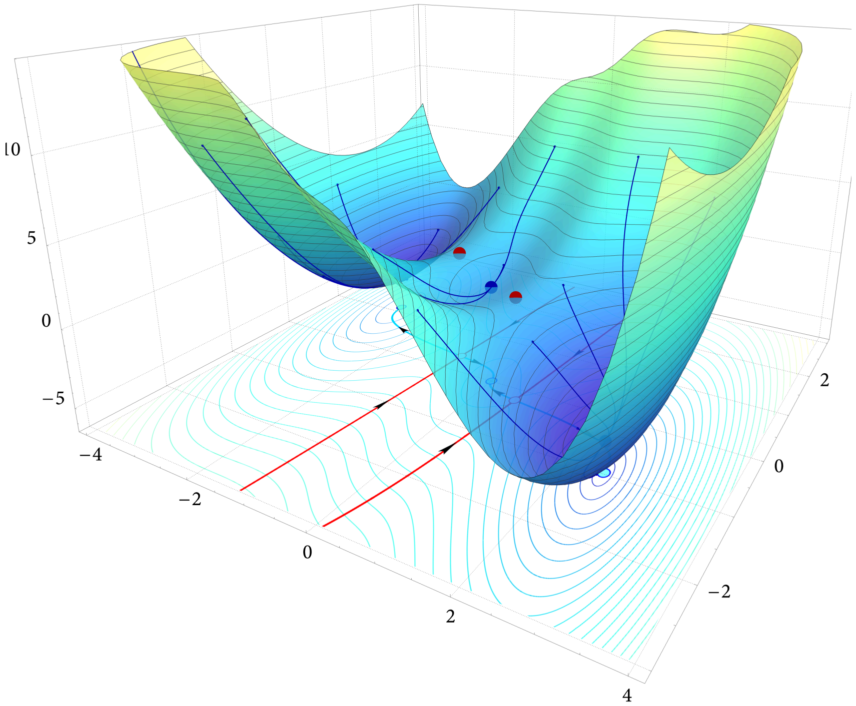
Power-law Gaussian bound:

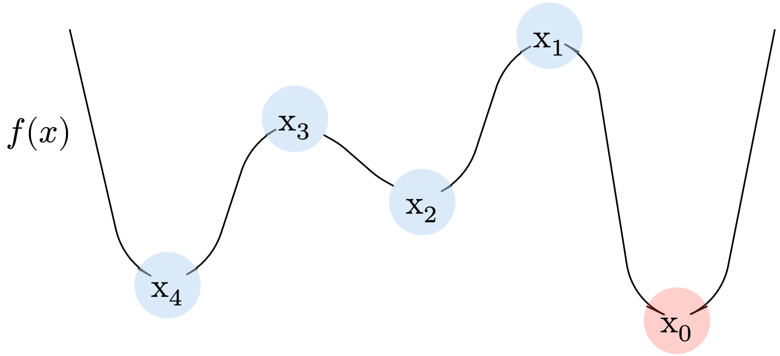
$$J(x) \leq \frac{2 \times \sharp\{\text{bad local minima}\} \times \{\text{log max. saddle - log min. bad local min.}\}}{\sigma^2}$$

$$J(x) \leq \frac{2 \times 2(\log f(x_1) - \log f(x_4))}{\sigma^2}$$



Example: Three Humps





$$f(x) = 2\frac{x_1^6}{13} + \frac{x_1^5}{8} - 91\frac{x_1^4}{64} - 24\frac{x_1^3}{48} + 42\frac{x_1^2}{16} + 5\frac{x_2^2}{4} + x_1 x_2$$

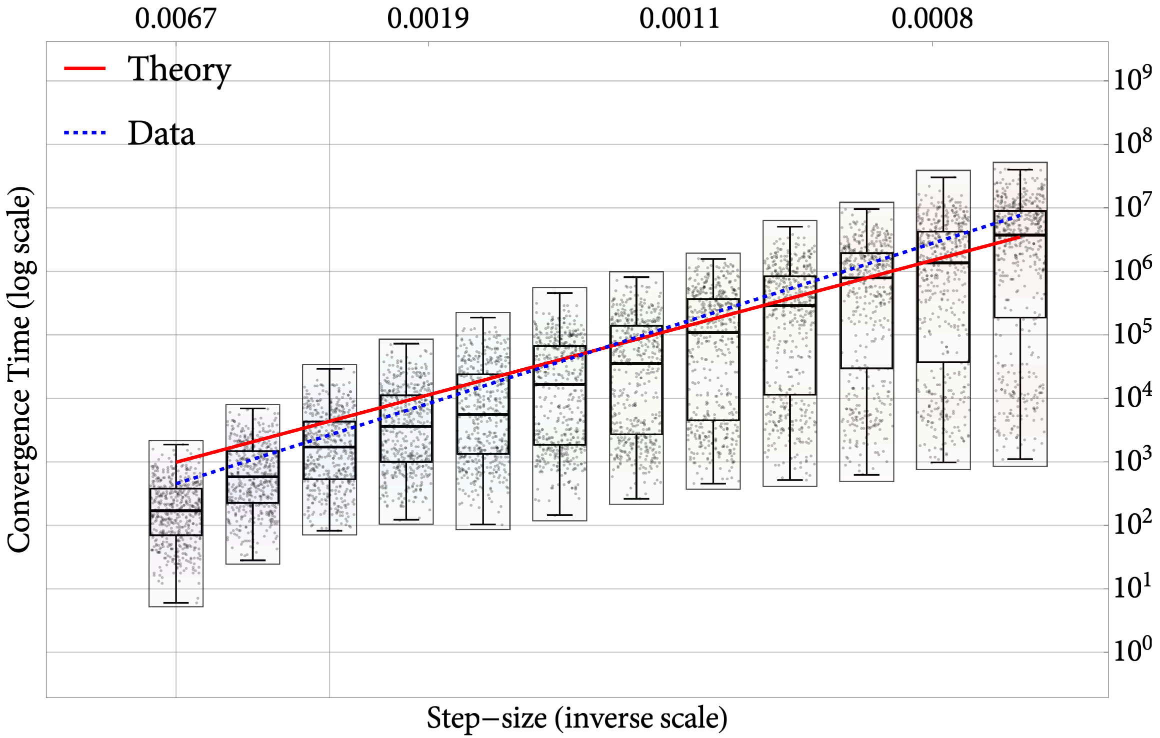
Three Humps: Simulation

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

we predict

$$J(x) = \frac{2(f(x_1) - f(x_4))}{\sigma^2}$$

$$\log \tau \approx \frac{2(f(x_1) - f(x_4))}{\sigma^2} \times \frac{1}{\eta}$$



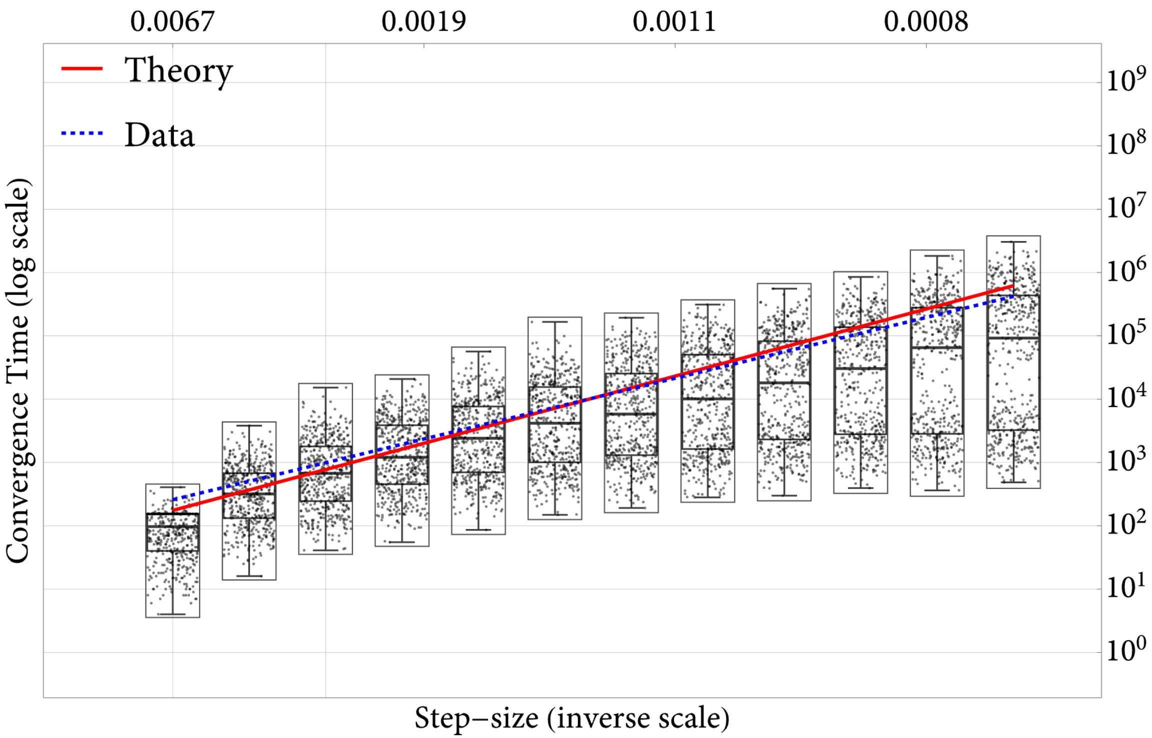
Three Humps: Simulation

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

we predict

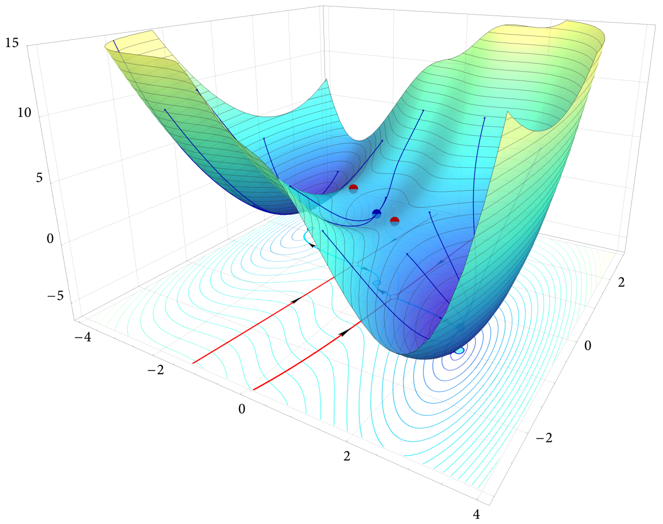
$$J(x) = \frac{2(\log f(x_1) - \log f(x_4))}{\sigma^2}$$

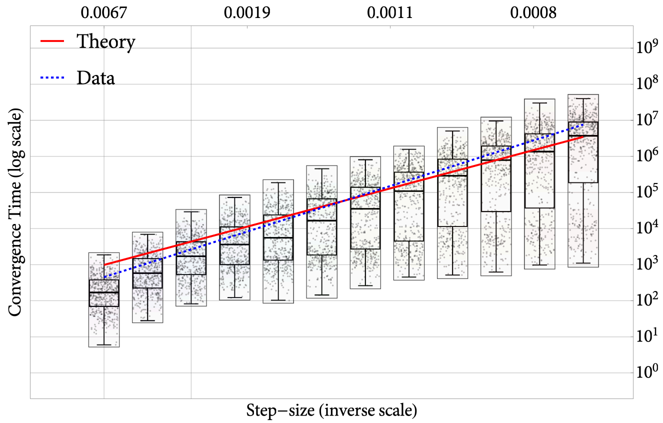
$$\log \tau \approx \frac{2(\log f(x_1) - \log f(x_4))}{\sigma^2} \times \frac{1}{\eta}$$



Partial Conclusion (second part)

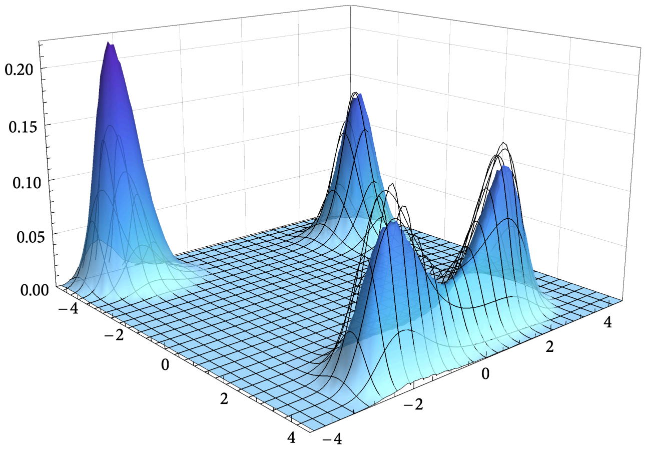
- We presented a characterization of the global convergence time of SGD
- The key quantity $J(x)$ captures the interplay between the loss landscape and the noise structure
- Built on our large deviation framework to analyze the long-term behavior of SGD

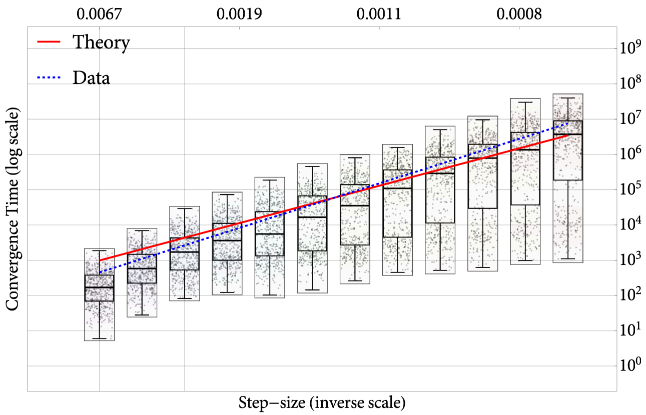




Conclusions and perspectives

- Provided answers to two fundamental questions about SGD in nonconvex problems,
- Intricate interplay between optimization, geometry, and noise in nonconvex learning problems.
- Answered these questions by developing a novel large deviation framework to analyze the long-term behavior of SGD.

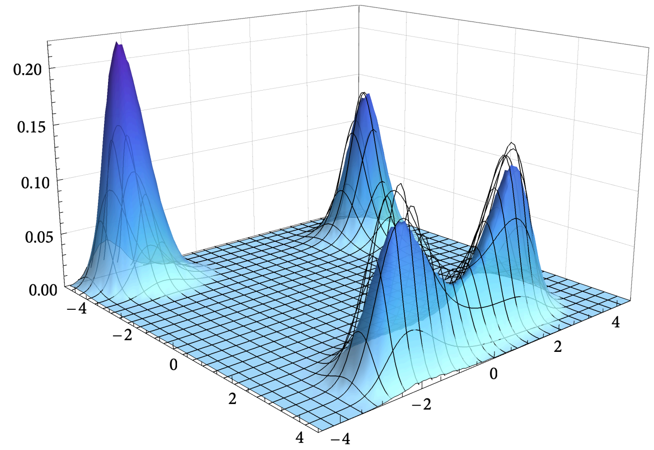


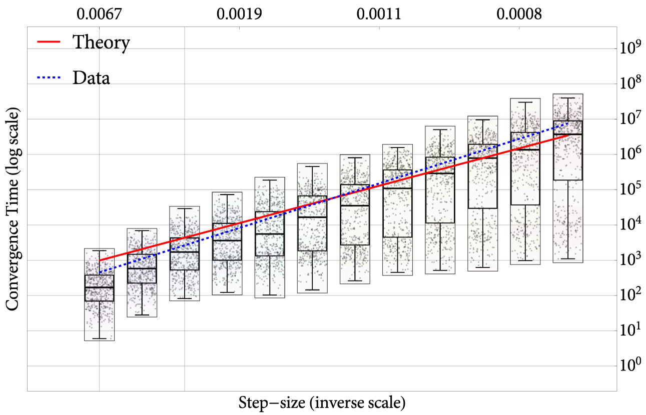


Conclusions and perspectives

- Provided answers to two fundamental questions about SGD in nonconvex problems,
- Intricate interplay between optimization, geometry, and noise in nonconvex learning problems.
- Answered these questions by developing a novel large deviation framework to analyze the long-term behavior of SGD.

- Coming next:
  - Analysis and design of adaptive methods
  - Understanding the implicit bias of SGD



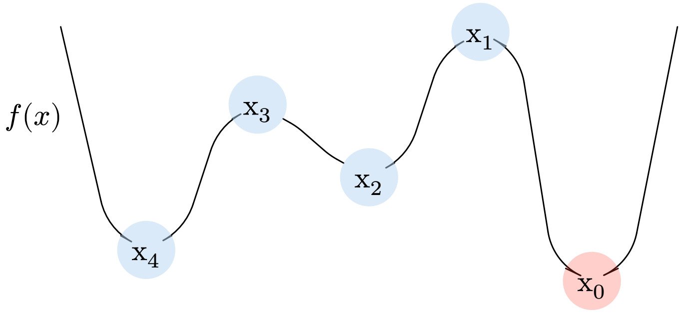


Generic bounds for energy

If $K_{i^*} = \operatorname{argmin} f$,

$$\sup_{j \neq i^*} \{E_{i^*} - E_j\} \leq \frac{2 \times \sharp\{\text{bad local minima}\} \times \{\text{max. saddle - min. bad local min.}\}}{\sigma^2_{\text{others}}} - \frac{2\{\text{depth of global min.}\}}{\sigma^2_{i^*}}$$

where $\sigma^2_{i^*}$ "upper bound" on the noise at K_{i^*} and σ^2_{others} "lower bound" on the noise at other components

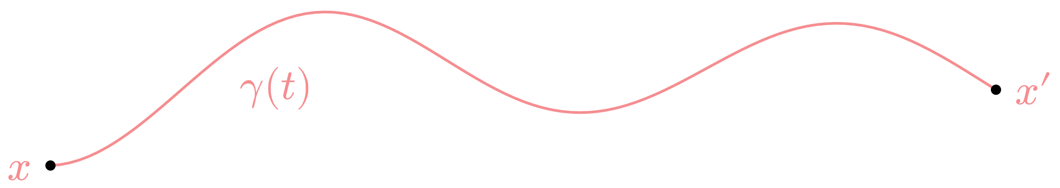


Quasi-potential

Following Kifer (1988), for any x, x'

$$B(x, x') = \inf\{\mathcal{S}_T[\gamma] \mid \gamma(0) = x, \gamma(T) = x', T \in \mathbb{N}\}$$

"$B(x, x')$ quantifies how probable a transition from x to x' is"



Key observations:

- If there is a trajectory of the gradient flow joining x and x', then $B(x, x') = 0$
- We can show:

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

Induced chain

Recall:

$$\mathrm{crit}(f) := \{x : \nabla f(x) = 0\} = \{K_1, K_2, ..., K_p\} \text{ with } K_i \text{ connected components}$$

(Conceptual) induced chain:

$z_n = i$ if the n-th visited component is K_i (up to a small neighborhood)

Goal: show that z_n captures the long-run behavior of SGD

Two key ingredients:

Ingredient 1 The behavior of SGD started at $x_0 \in K_i$ depends only on i.

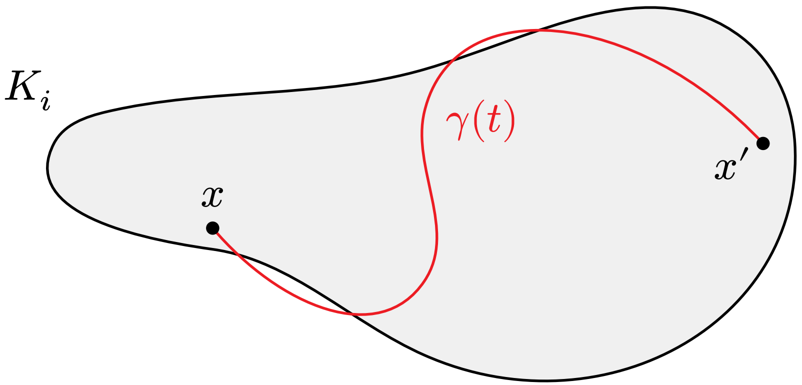
Ingredient 2 SGD spends most of its time it near $\mathrm{crit}(f)$.

Ingredient 1

Equivalence relation:

for $x, x' \in \mathrm{crit}(f)$, $\quad x \sim x' \Leftrightarrow B(x, x') = B(x', x) = 0$

> ***Proposition:***
> *if the K_i are connected by smooth arcs, the equivalence classes of $\sim$ are exactly $K_1, \ldots, K_p$*

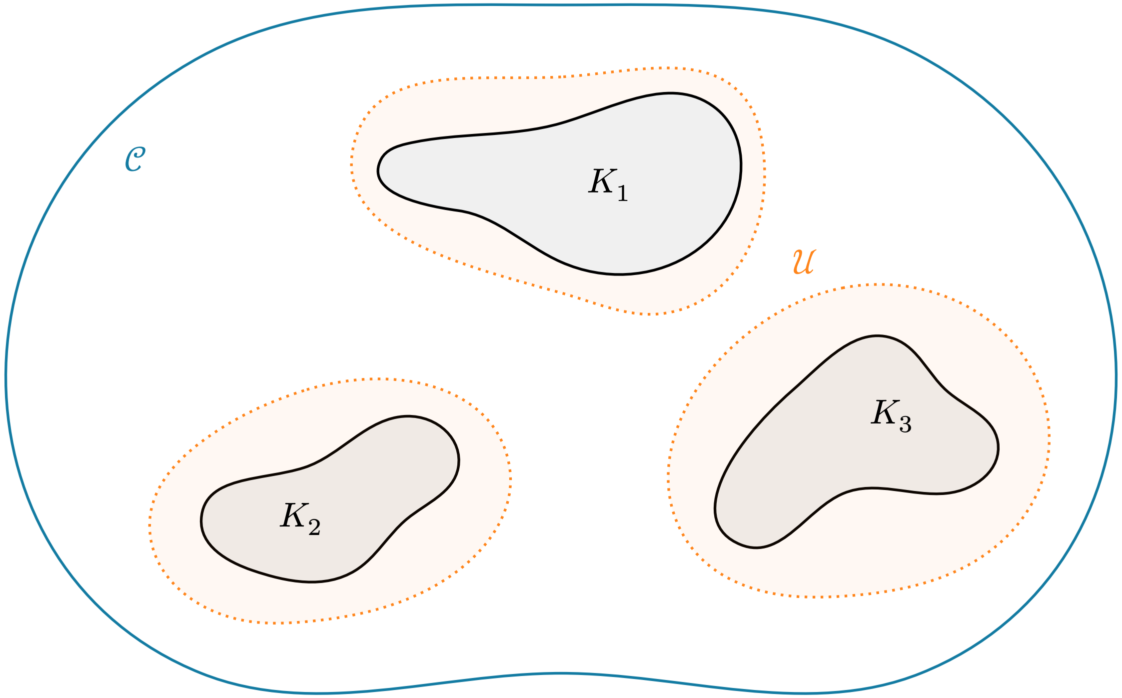


"Behaviour of SGD started at $x \approx$ Behaviour of SGD started at x'"

Ingredient 2

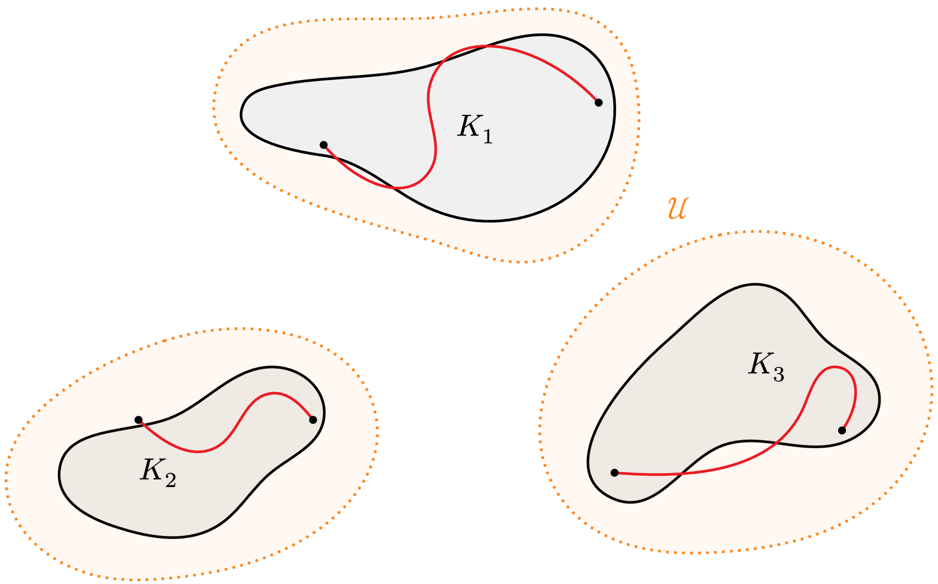
Proposition: *given* $\mathrm{crit}(f) \subset \mathcal{U} \subset \mathcal{C}$ *with* $\mathcal{U}$ *open,* $\mathcal{C}$ *compact, for* $\eta > 0$ *small enough,*

$$\forall x \in \mathcal{C}, \qquad \mathbb{P}\Big(\textit{SGD started at } x \textit{ reaches } \mathcal{U} \textit{ in} \geq n \textit{ steps}\Big) \leq e^{-\Omega\left(\frac{n}{\eta}\right)}$$



Induced chain

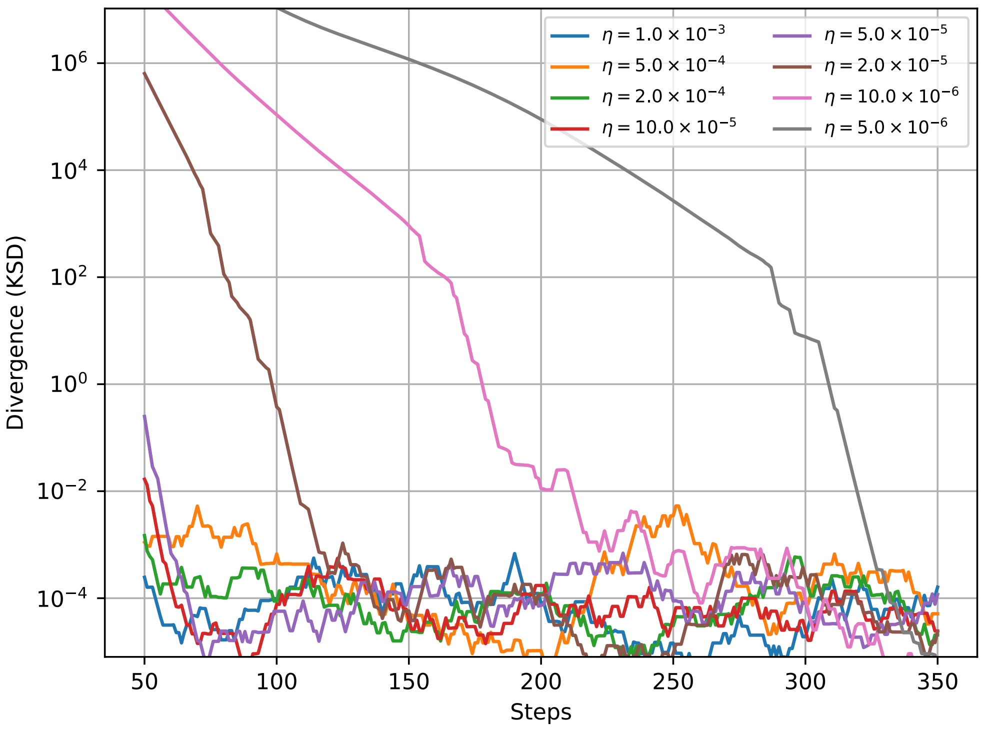
(Conceptual) induced chain:
$z_n = i$ if the n-th visited component is K_i (up to a small neighborhood)



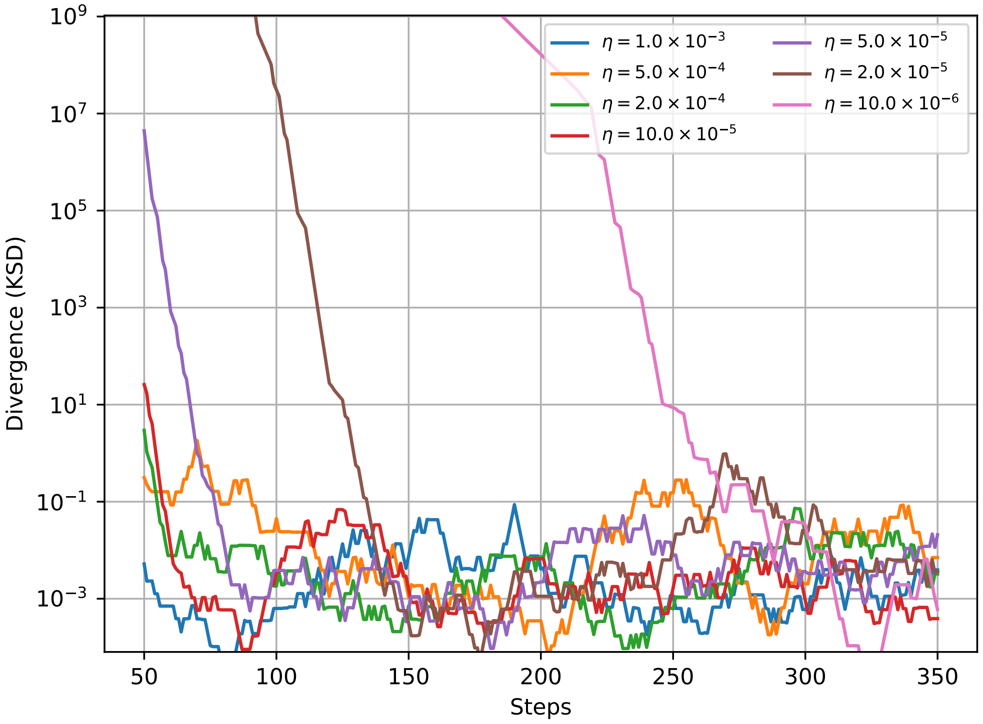
Ingredients 1 + 2 imply

The induced chain z_n captures the long-run behavior of SGD

Example: Back to Himmelblau



Divergence between iterates and Gibbs
$Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I_d)$



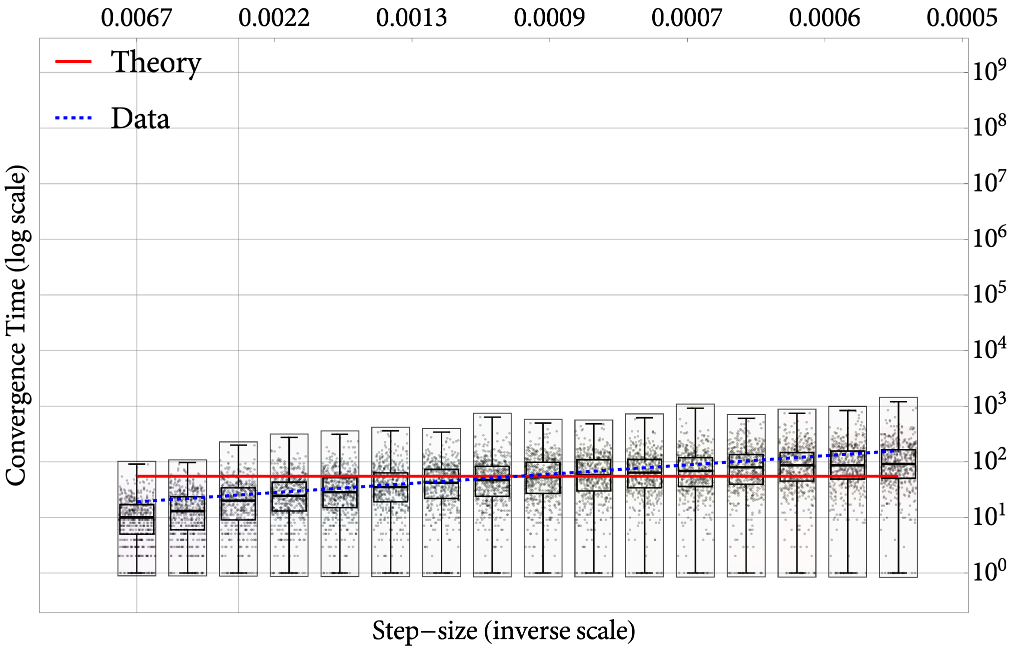
Divergence between iterates and Power-law
$Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 f(x) I_d)$

No spurious local minima: Simulation

Consider a non-convex function (with maxima and saddles) but no spurious local minima:

We predict $E = 0$ and thus

$$\log \tau = \text{cst}$$



Attempt with SDE

1. Approximate SGD by an SDE:

$$dX_t = -\nabla f(X_t)dt + \sqrt{\eta\Sigma(X_t)}dW_t$$

2. Combine with the exponential convergence of the SDE to its invariant measure

$$X_t \underset{t\to\infty}{\to} \mu_\infty$$

But: convergence speed of the SDE is not fast enough to compensate for the approximation error!