What is the Long-Run Behaviour of SGD?

A Large Deviation Analysis

Séminaire Probailités / Statistiques, Université de Nice

December 3, 2023

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

Deep learning

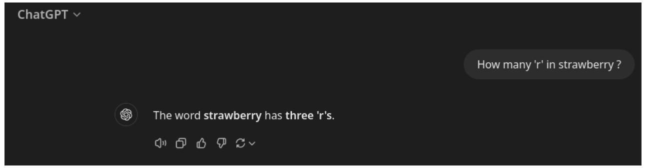


Image credit: Meta AI

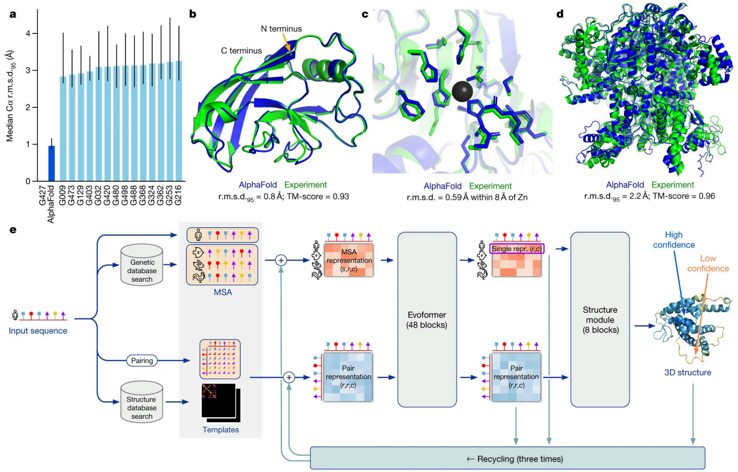


Image credit: DeepMind

Training: minimizing the loss of the model on data

Problem of interest (finite-sum)

For $f : \mathbb{R}^d \to \mathbb{R}$ smooth

$$\underset{x \in \mathbb{R}^d}{\text{minimize}}\, f(x) \qquad \text{where} \qquad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

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

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

Problem of interest

For $f : \mathbb{R}^d \to \mathbb{R}$ smooth

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

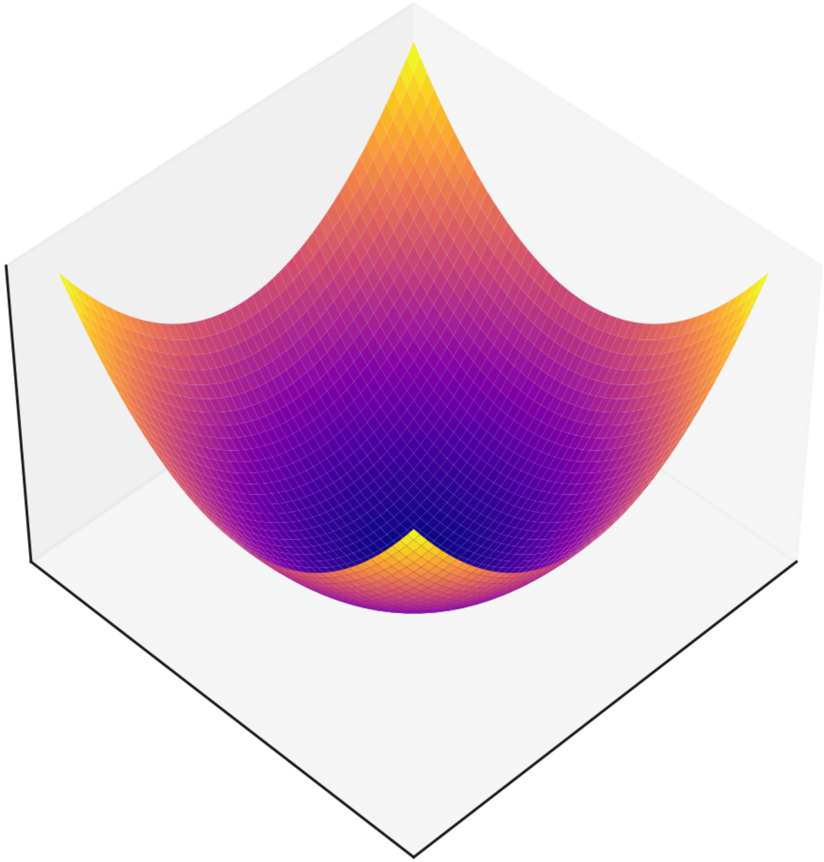
Stochastic Gradient Descent (SGD): with *constant* step-size $\eta > 0$

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

step-size   zero-mean noise

Q: What is the asymptotic behaviour of SGD?

Convex loss



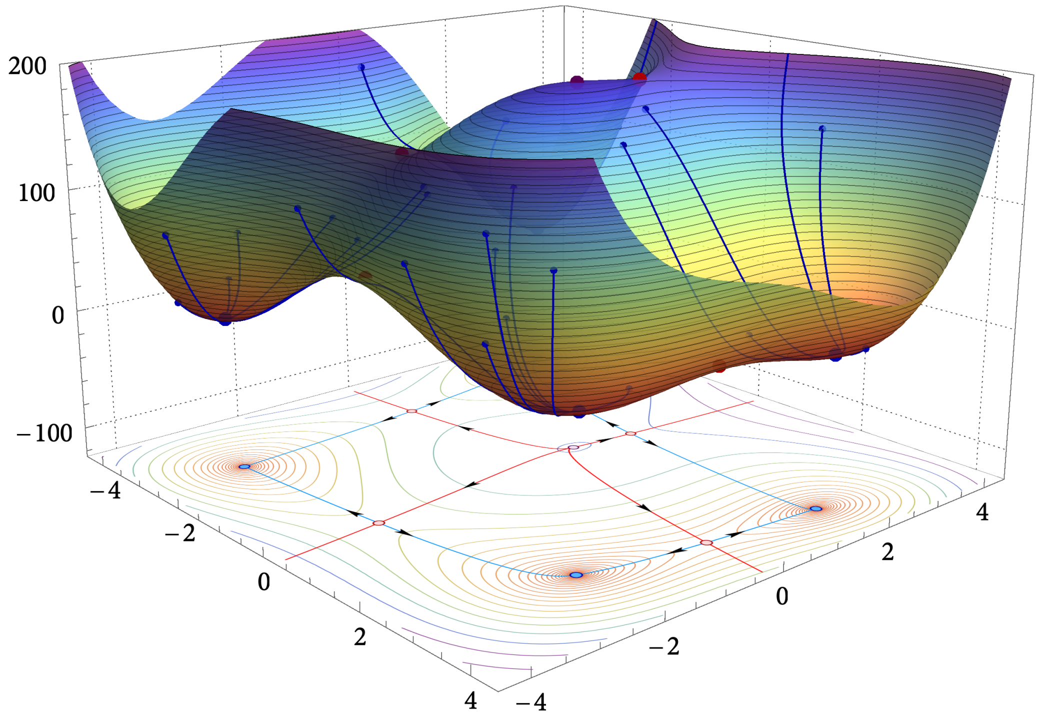
Nonconvex loss!

Image credit: losslandscape.com

Training of deep neural networks = SGD on a nonconvex loss function

Himmelblau function

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



Himmelblau function

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:

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

  Convergence to local minima (Bertsekas & Tsitsiklis, 2000) but no information about which one.

- Sampling (MCMC, Langevin):

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

  Scaling of the noise differs from SGD ⇒ analysis does not carry over

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

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

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

What is known?

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

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

SGD with constant step-size:

- f strongly convex: SGD converges near the minimizer
- f convex: average of SGD iterates (almost) optimal
- 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)

Q: Which critical points (and which local minima) are visited the most in the long run?

New approach: large deviations

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

Published and presented at ICML 2024, Vienna, Austria

Outline:

1. Informal result
2. Less informal overview of the approach

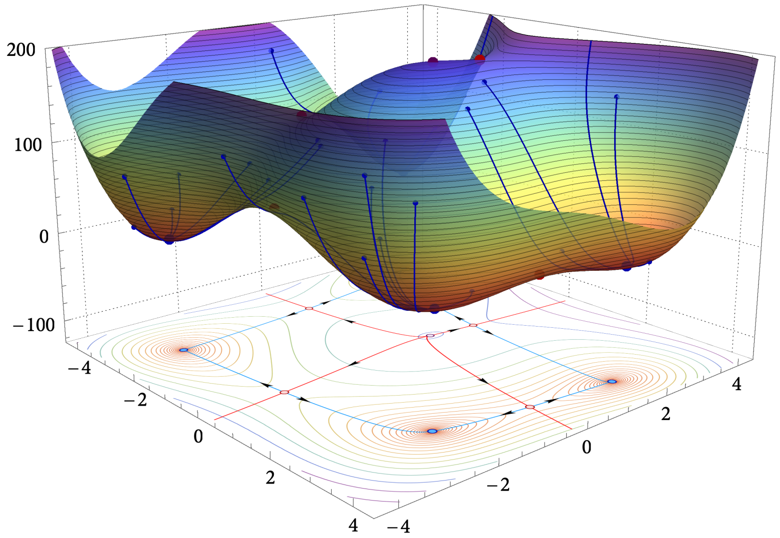
On the objective function f

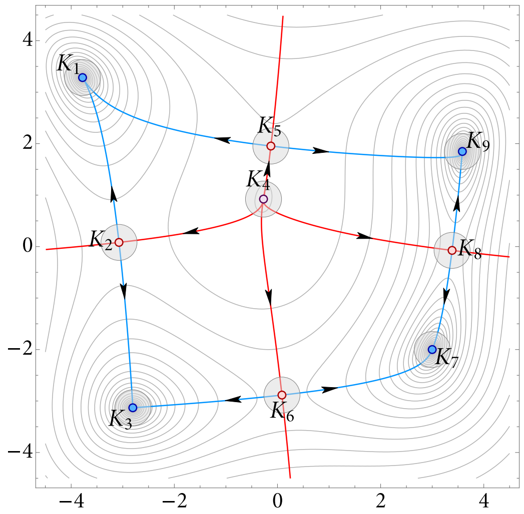
Regularity assumption:

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

where K_i connected components (compact)

Himmelblau function





Asymptotic behaviour

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

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

Q: Where do invariant measures of SGD concentrate?

Main results (informal)

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

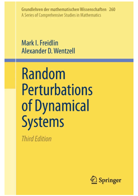
Challenges and techniques

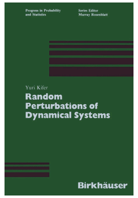
- No known approach to analyze the asymptotic distribution of SGD on non-convex problems
- We leverage large deviation theory and the theory of random perturbations of dynamical systems,
  $\to$ Estimate the probability of rare events, such as SGD escaping a local minima
- We adapt the theory of random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

Objective assumptions:
- f β-smooth, i.e. ∇f is β-Lipschitz
- 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:

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

Example (Finite-sum):

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

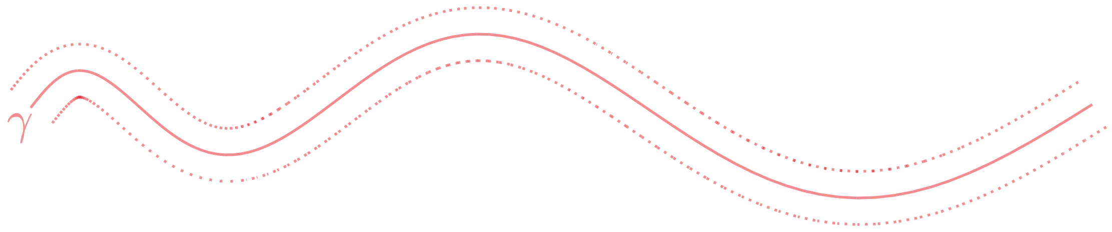
SGD :

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

with $Z(x;\omega) = \nabla f_\omega(x) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x)$

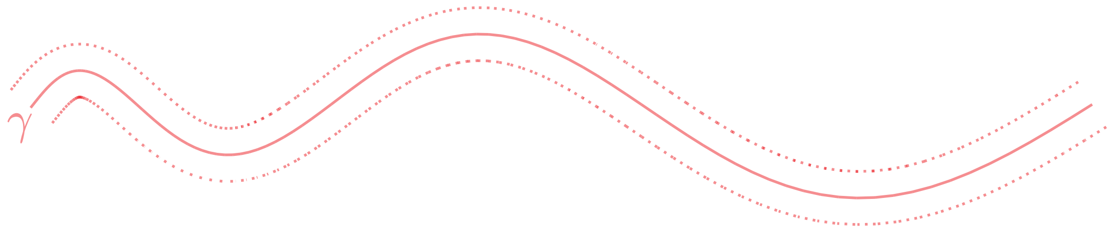
Large deviations for SGD

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



Large deviations for SGD

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



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

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

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:

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

Key observations:

- _____ iff $\mathcal{S}_T[\gamma] = 0$
- The farther γ is from being a gradient flow, the _____ $\mathcal{S}_T[\gamma]$
- And, as a consequence, the _____ the probability of SGD following γ

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

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: $\dot{\gamma}_t = -\nabla f(\gamma_t)$ 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 _____ the probability of SGD following γ

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:

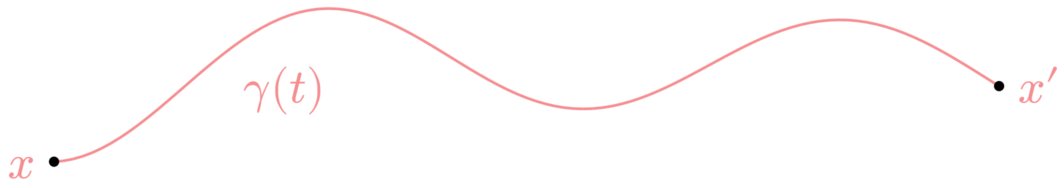
- γ is a trajectory of a gradient flow: $\dot{\gamma}_t = -\nabla f(\gamma_t)$ iff $\mathcal{S}_T[\gamma] = 0$
- The farther γ is from being a gradient flow, the larger $\mathcal{S}_T[\gamma]$
- And, as a consequence, the smaller the probability of SGD following γ

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 $B(x,x') = 0$
- It holds:

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

Induced chain

Recall:

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

(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 behaviour of SGD started at $x_0 \in K_i$ depends only on i.

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

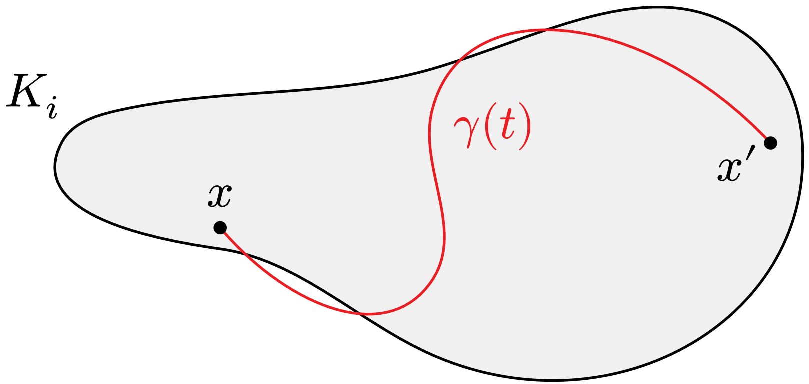
Ingredient 1

Equivalence relation:

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

Proposition:

if the K_i are connected by smooth arcs, the equivalence classes of $\sim$ are exactly $K_1, \dots, K_p$

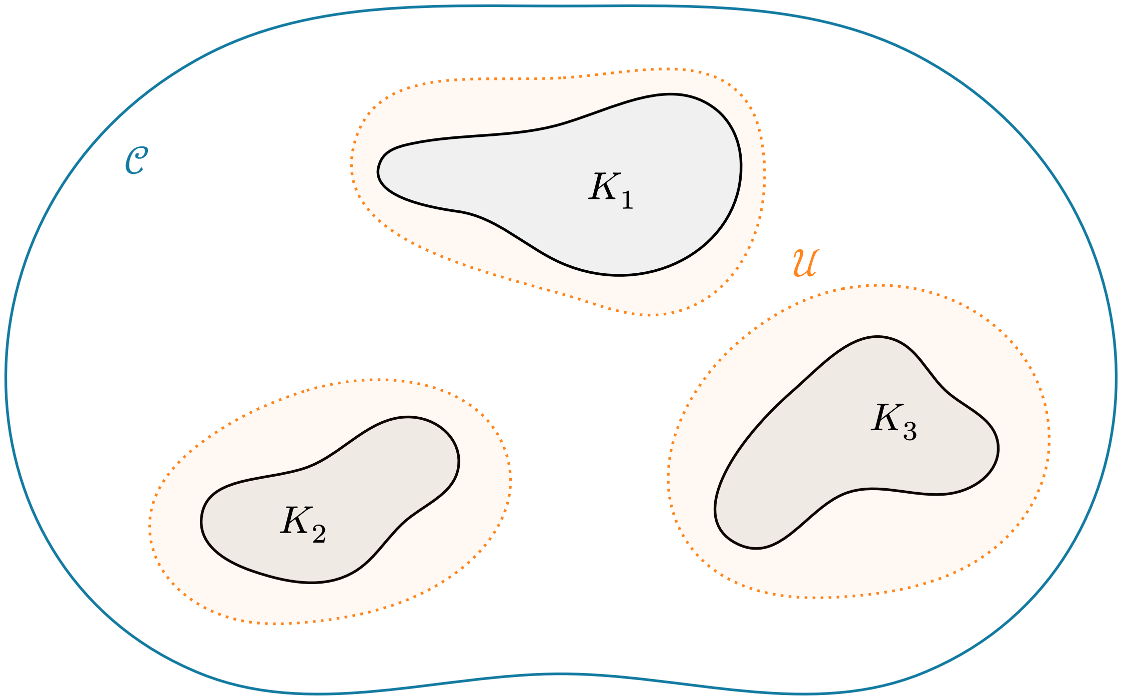


"Behaviour of SGD started at $x \approx$ Behaviour of SGD started at x'"

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,*

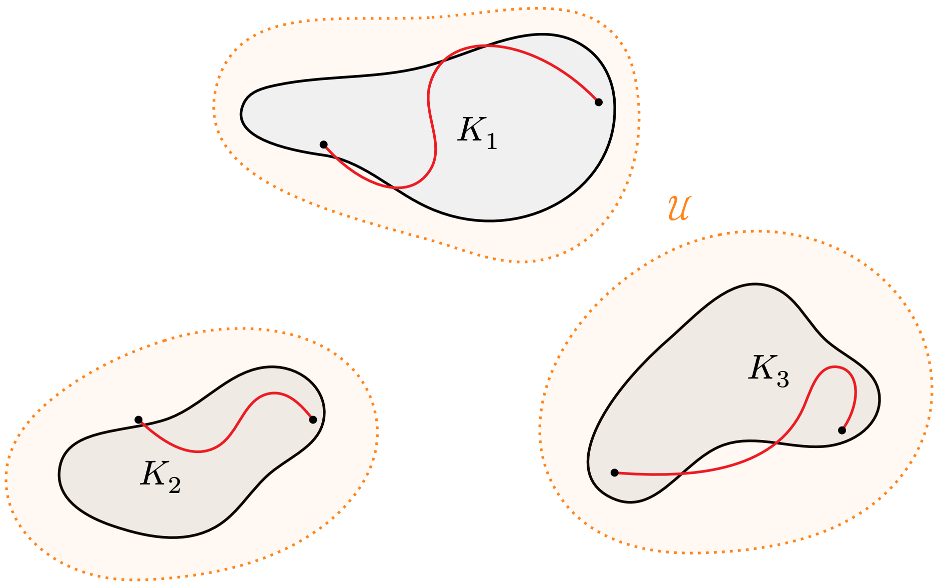
$$\forall x \in \mathcal{C}, \qquad \mathbb{P}\Big(\text{SGD started at } x \text{ reaches } \mathcal{U} \text{ in} \geq n \text{ steps}\Big) \leq e^{-\Omega\left(\frac{n}{\eta}\right)}$$



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

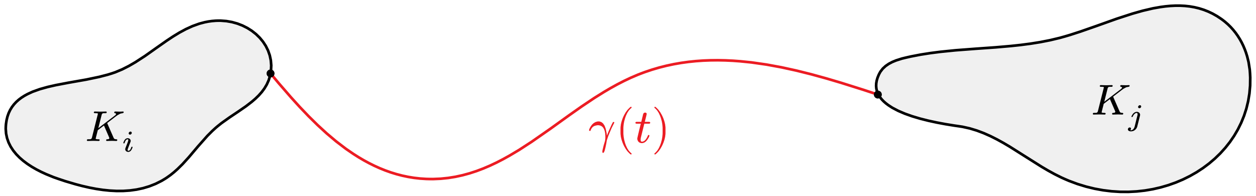
The induced chain z_n captures the long-run behavior of SGD

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:

$$B_{i,j} = \inf\{B(x_i, x_j) \mid x_i \in K_i, x_j \in K_j\} = \inf\{\mathcal{S}_T[\gamma] \mid \gamma(0) = K_i, \gamma(T) = K_j, T \in \mathbb{N}\}$$

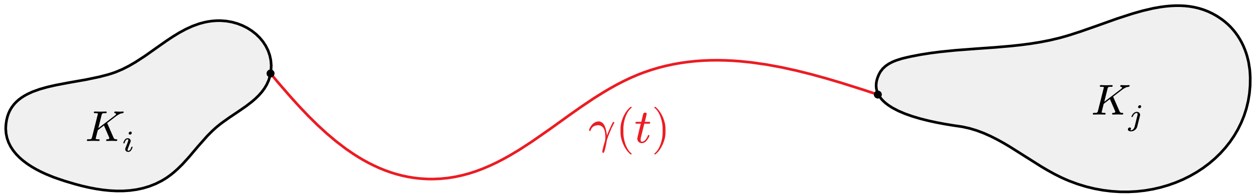


Transition between critical points

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:

$$B_{i,j} = \inf\{B(x_i, x_j) \mid x_i \in K_i, x_j \in K_j\} = \inf\{\mathcal{S}_T[\gamma] \mid \gamma(0) = K_i, \gamma(T) = K_j, T \in \mathbb{N}\}$$



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

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

Transition graph

Now, study z_n as a Markov chain on $\{1, ..., p\}$ with $\mathbb{P}(z_{n+1} = j \mid z_n = i) \approx \exp\left(-\frac{B_{i,j}}{\eta}\right)$

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 of K_i:

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

Lemma *(very informal)**:** the invariant measure of z_n is, for $\eta > 0$ small enough,*

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

Main results (more formal)

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* $E_j < E_i$*:*

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

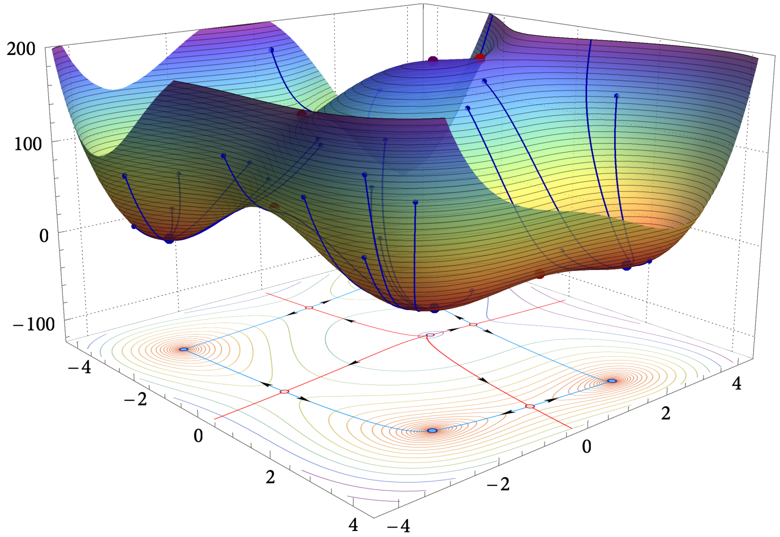
4. ***Concentration on ground states:*** *given* $\mathcal{U}_0$ *neighborhood of the ground states* $K_0 = \operatorname{argmin}_i E_i$

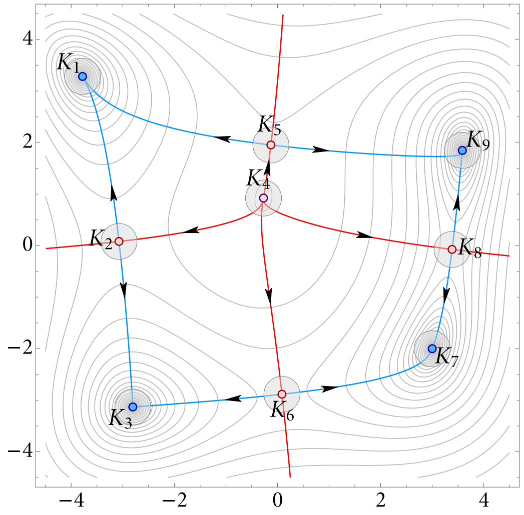
$$\mu_\infty(\mathcal{U}_0) \geq 1 - e^{-\frac{\lambda_0}{\eta}}, \qquad \text{for some } \lambda_0 > 0$$

Example: Gaussian noise

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

Himmelblau function



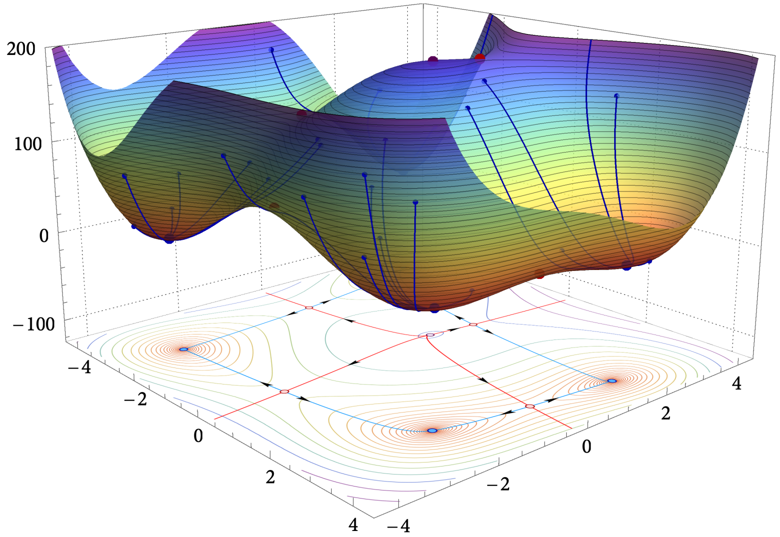


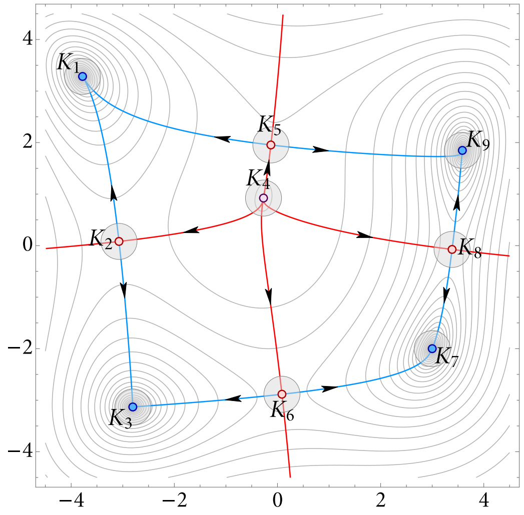
$$B_{5,1} = 0; \qquad B_{1,5} = \frac{2(f(K_5) - f(K_1))}{\sigma^2}$$

Example: Gaussian noise

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

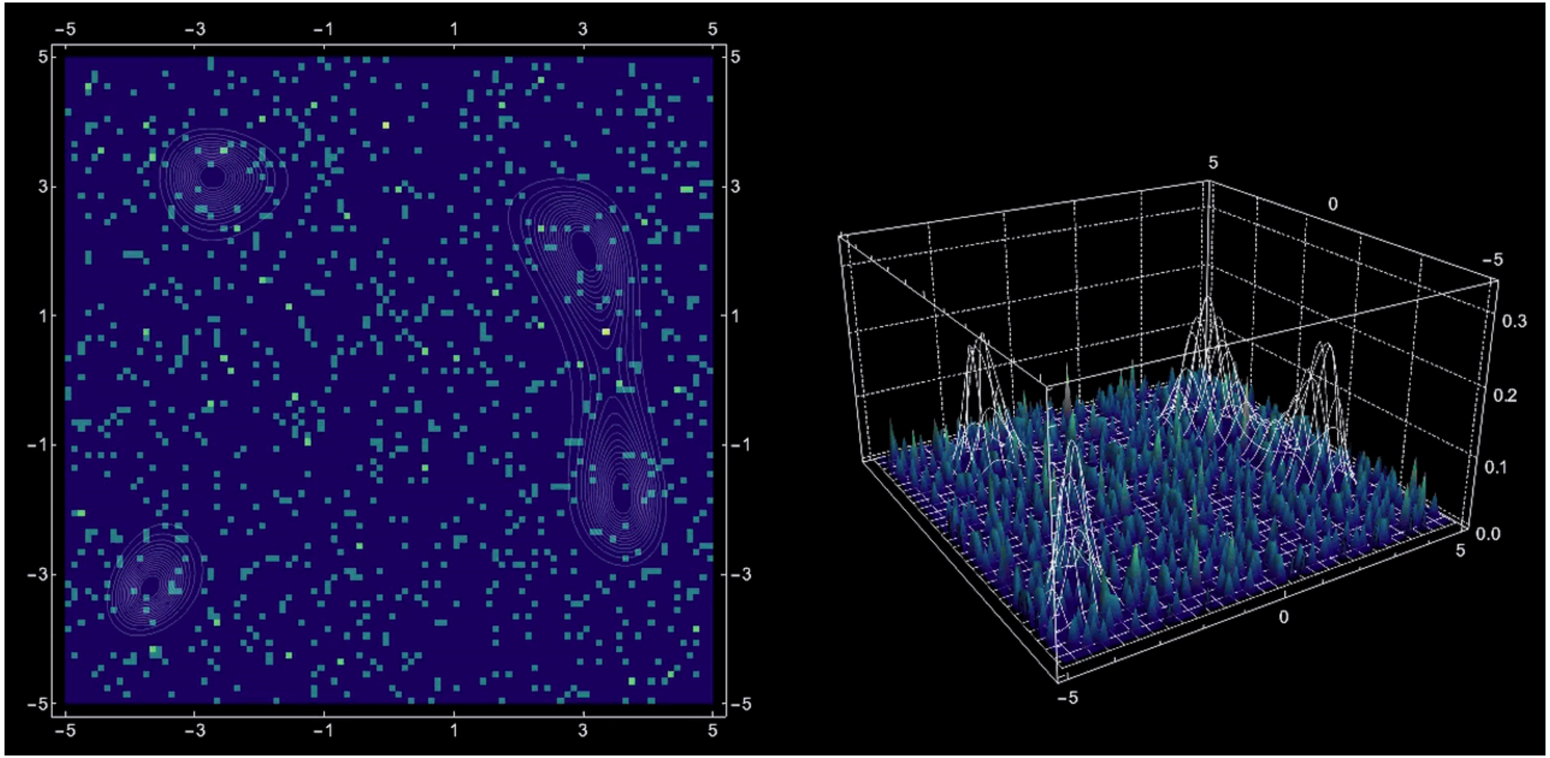
Himmelblau function





$$E_i = \frac{2f(x_i)}{\sigma^2} \text{ for any } x_i \in K_i$$

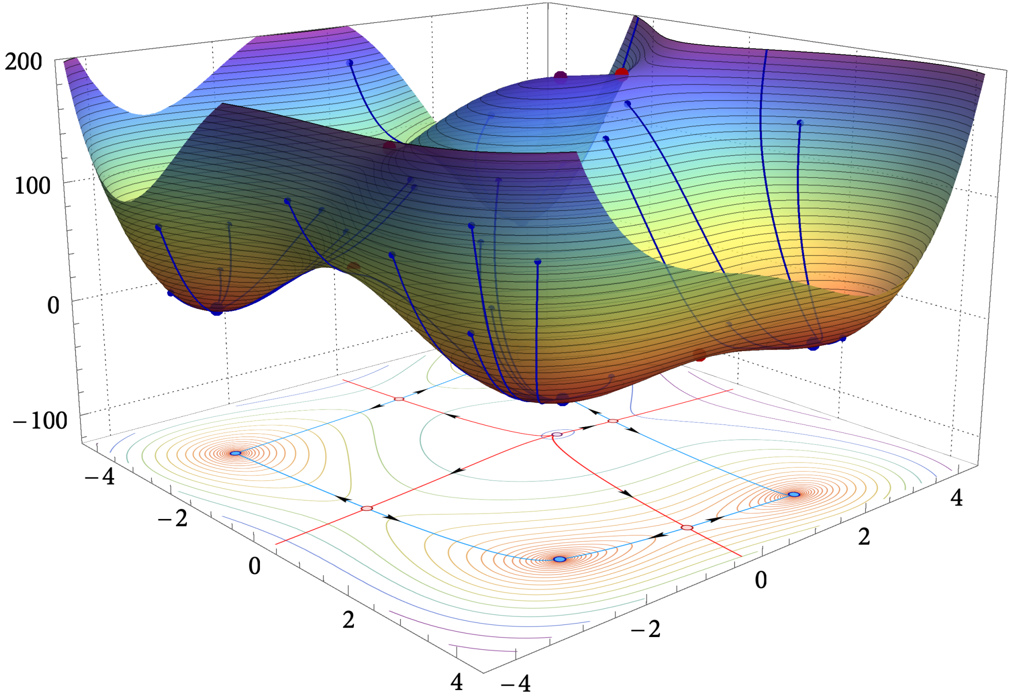
Example: Gaussian noise



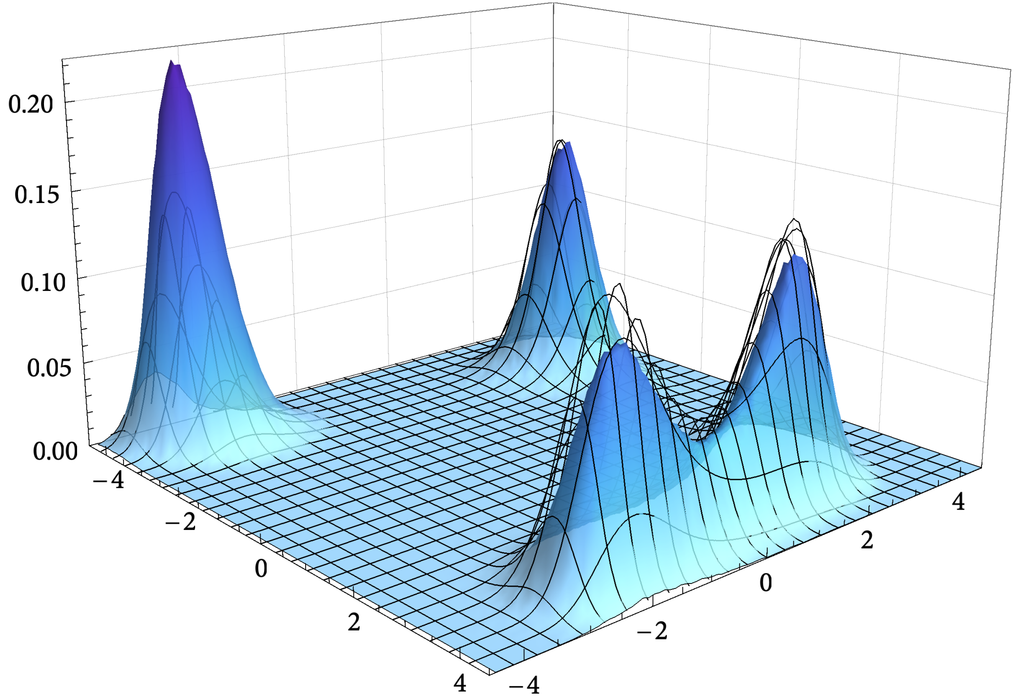
Evolution of the distribution of the iterates of SGD

Example: Gaussian noise

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



Himmelblau function



Simulation vs prediction of the invariant measure

Conclusion

- We introduce a theory of large deviation for SGD in nonconvex problems.
- We demonstrate its potential by characterizing the asymptotic distribution of SGD.

arXiv:2406.09241

Image credit: losslandscape.com

Conclusion

- We introduce a theory of large deviation for SGD in nonconvex problems.
- We demonstrate its potential by characterizing the asymptotic distribution of SGD.

- Coming next:
  - Adaptive methods
  - Explicit bounds and time to convergence
  - Link to the geometry of the loss landscape of neural networks

arXiv:2406.09241

Image credit: losslandscape.com