What is the Long-Run Distribution of SGD?

A Large Deviation Analysis

October 2024

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

Problem of interest

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

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

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

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

Q: What is the asymptotic distribution of SGD?

What is known?

- f strongly convex: SGD converges to (almost) the minimizer
- f convex: average of SGD iterates (almost) converges to minimizers
- f nonconvex:
  - In average, close to critical points (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 (Lee et al., 2016, 2017)

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

New approach: large deviations

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

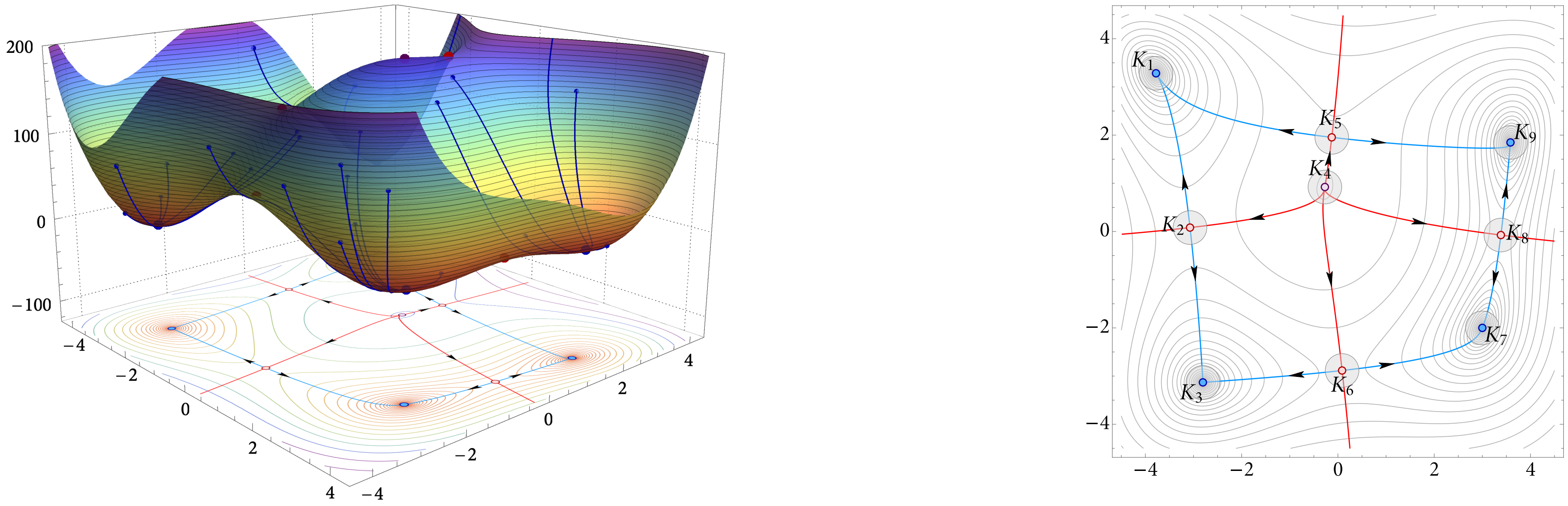
Outline:

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

On the objective function f

Simplified framework:

$$\operatorname{crit}(f) := \{x : \nabla f(x) = 0\} = \{c_1, c_2, \ldots, c_p\}$$



Himmelblau function

In the paper: connected components $K_1, K_2, \ldots, K_p$

Asymptotic distribution

Invariant measure: probability measure μ_∞ such that

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

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

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

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

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,

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

4. **Ground state concentration:** there is i_0 such that

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

Challenges and techniques

- No known approach to analyze the asymptotic distribution of SGD on non-convex problems
  e.g. SDE approximations only valid on finite time horizons
- We leverage large deviation theory and the theory of random dynamical systems,
  $\rightarrow$ Estimate the probability of rare events, such as SGD escaping a local minima
- We adapt the theory of Freidlin & Wentzell (1998); Kifer (1988) to SGD with two main challenges:
  a) Lack of compactness
  b) Realistic noise models (finite sum)
  $\rightarrow$ 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

Objective assumptions: f is coercive and β-smooth

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:
$$\log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right] \leq \tfrac{\sigma^2}{2}\|v\|^2$$
- SNR high enough:
$$\liminf_{\|x\|\to\infty} \tfrac{\|\nabla f(x)\|^2}{\sigma^2} \quad \text{larger than some constant}$$

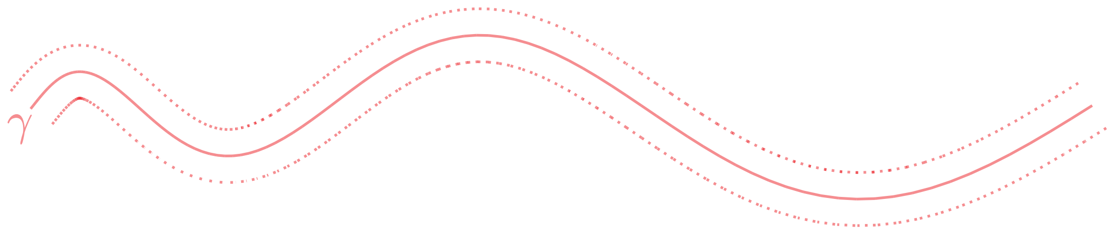
Example (Finite-sum):

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

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

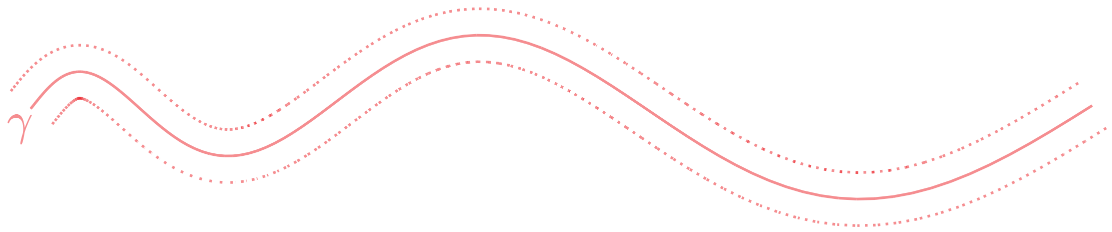
Large deviations for SGD

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, $\mathbb{P}(\text{SGD} \approx \gamma) = ?$



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

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

with

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

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

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:

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

LDP in the Gaussian case

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

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:

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

LDP in the Gaussian case

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

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:

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

LDP in the Gaussian case

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

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:

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

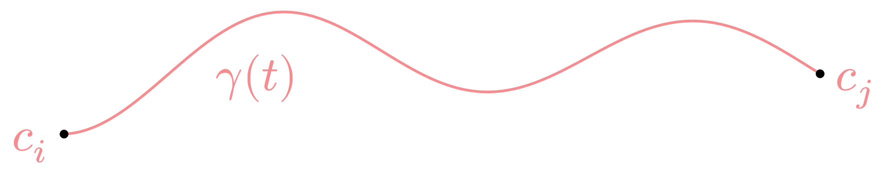
Transition between critical points

Given c_i, c_j critical points, what is

$$\mathbb{P}\big(\text{SGD transitions from } c_i \text{ to } c_j\big)?$$

Involves the transition cost:

$$B_{i,j} = \inf\big\{\mathcal{S}_T[\gamma] \mid \gamma(0) = c_i, \gamma(T) = c_j, T \in \mathbb{N}\big\}$$



Transition graph

Proposition: *Transition probability from c_i to c_j:*

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

where $B_{i,j}$ transition cost

$$B_{i,j} = \inf\big\{\mathcal{S}_T[\gamma] \mid \gamma(0) \in c_i, \gamma(T) \in c_j, T \in \mathbb{N}\big\}$$

Technical assumption: $B_{i,j} < +\infty$ for all c_i, c_j

Transition graph: *complete graph on $\{c_1, ..., c_p\}$ with weights $B_{i,j}$ on $i \to j$*

Energy *of c_i:*

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

Main results (more formal)

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

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

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

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

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

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

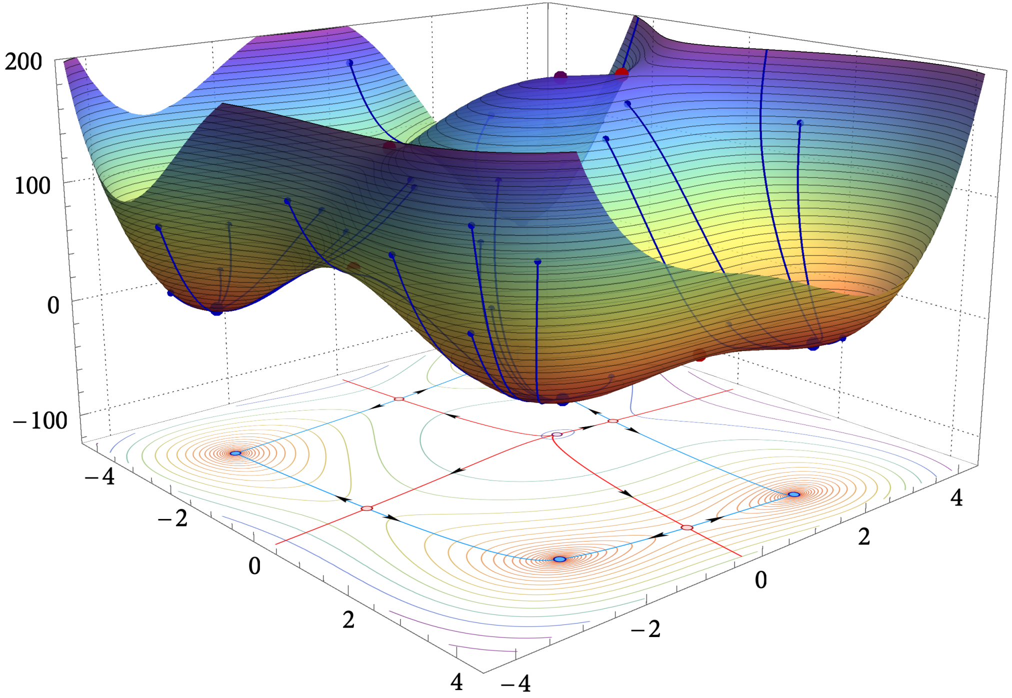
$$\frac{\mu_\infty(\mathcal{U}_i)}{\mu_\infty(\mathcal{U}_j)} \leq e^{-\frac{\lambda_{i,j}}{\eta}} \qquad \text{for some } \lambda_{i,j} > 0$$

4. *Concentration on ground states: given* $\mathcal{U}_0$ *neighborhood of the ground states* $c_0 = \mathrm{argmin}_i\, E_i$*,*

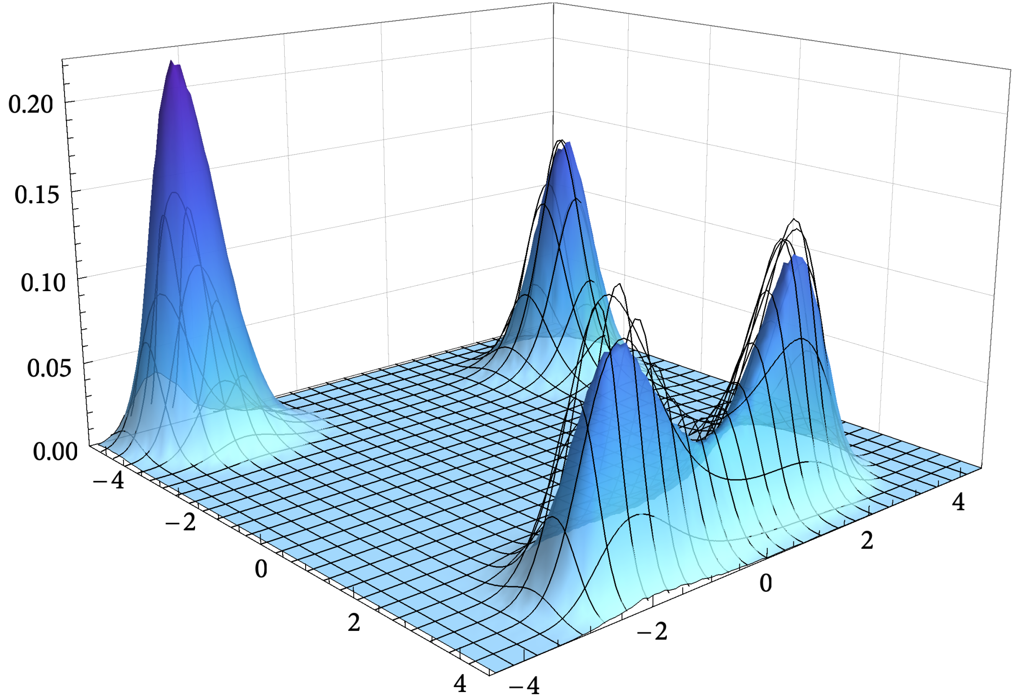
$$\mu_\infty(\mathcal{U}_0) \geq 1 - e^{-\frac{\lambda_0}{\eta}}, \qquad \text{for some } \lambda_0 > 0$$

Example: Gaussian noise

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



Himmelblau function



Simulation vs prediction of the invariant measure

Conclusion: a first step towards understanding nonconvex problems

1. We introduce a theory of large deviation for SGD in nonconvex problems.
2. We demonstrate its potential by characterizing the asymptotic distribution of SGD.
3. Coming next:
   - Explicit bounds
   - Time to convergence (reach some particular minima, converge to the invariant measure)
   - Link to the geometry of the loss landscape of neural networks

Image credit: losslandscape.com