What is the Long-Run Distribution of SGD? A Large Deviation Analysis

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

arXiv:2406.09241

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

Problem of interest

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

Stochastic Gradient Descent (SGD):

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

Basic assumptions:
- f is β-smooth: $\|\nabla f(x) - \nabla f(x')\| \le \beta\|x - x'\|$ for all x, x'
- f is coercive: $\lim_{\|x\|\to\infty} f(x) = +\infty$

Critical points

SGD spends most of its time on average near critical points $\operatorname{crit}(f) = \{x \in \mathbb{R}^d \mid \nabla f(x) = 0\}$

Q: Which critical points are more likely to be visited by SGD and by how much?

Regularity assumption:

$$\operatorname{crit}(f) = \bigcup_{i=1}^{p} K_i, \quad \text{where } K_i \text{ (smoothly) connected components}$$

Invariant measure

Invariant measure: probability measure μ_∞ on $\mathbb{R}^d$ such that

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

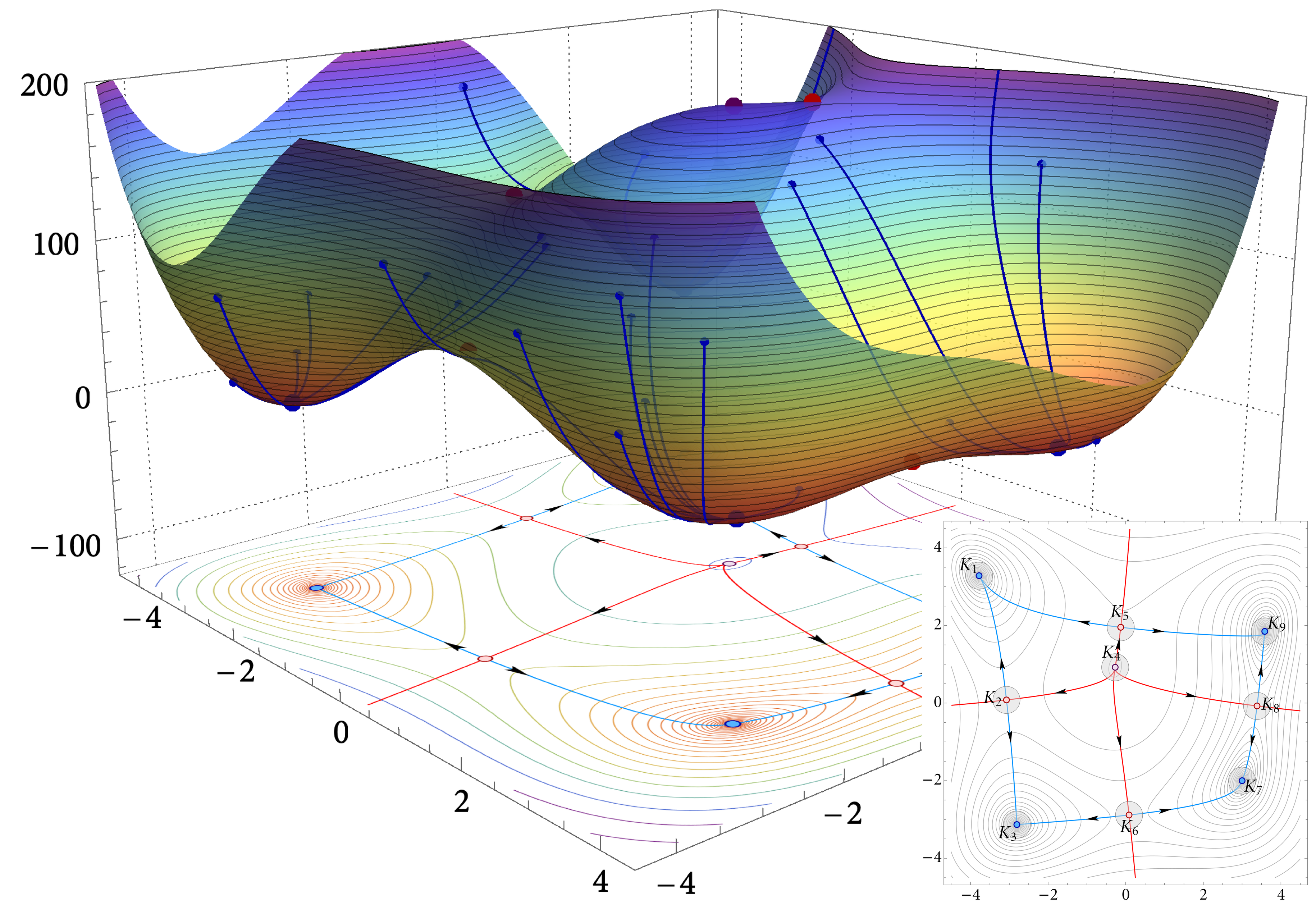
→ weak-⋆ limit points of the mean occupation measures of the iterates of SGD:

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

Q: Where do invariant measures of SGD concentrate?

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

Example (Finite-sum):
$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \frac{\lambda}{2}\|x\|^2$ with f_i Lipschitz and smooth; $Z(x;\omega) = \nabla f_\omega(x) - \nabla f(x)$



Main results:

1. **Concentration near critical points:** The iterates of SGD are exponentially concentrated near the critical points of f
2. **Saddle-point avoidance:** Non-minimizing critical points are exponentially less likely to be observed than local minimizers
3. **Boltzmann-Gibbs distribution:** The probability of observing a critical point is exponentially proportional to its energy (not its value)
4. **Ground state concentration:** The iterates of SGD are exponentially more likely to be observed near its ground state (set of minimum energy)

Challenges and techniques:

- No known approach to analyze the asymptotic distribution of SGD in 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,
  → 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)
  → 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

Noise statistics:

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

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

Lemma (Large deviations for SGD)

SGD admits a large deviation principle as $\eta \to 0$ with action functional

$$\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}(\gamma_t, \dot\gamma_t)dt \ \textit{ for any } \gamma : [0,T] \longrightarrow \mathbb{R}^d \textit{ abs. continuous on } [0,T]$$

This means: for $\eta > 0$ small enough, for any trajectory $\gamma : [0,T] \to \mathbb{R}^d$,

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

Example (Gaussian noise): $Z(x;\omega) \sim N(0, \sigma^2 I_d)$, $\mathcal{H}(x,v) = \frac{\sigma^2}{2}\|v\|^2$ and $\mathcal{L}(x,v) = \frac{\|v + \nabla f(x)\|^2}{2\sigma^2}$

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

Proposition (Transition probability)

Transition probability from K_i to K_j:

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

where $B_{i,j}$ transition cost

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

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

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

Energy *of K_i:*

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

Theorem (Invariant measure of SGD)

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

1. *Concentration on $\operatorname{crit}(f)$: there is some $c > 0$ s.t.*
$$\mu_\infty\left(\textstyle\bigcup_{i=1}^{p} \mathcal{U}_i\right) \ge 1 - e^{-\frac{c}{\eta}}, \qquad \textit{for some } c > 0$$
2. *Boltzmann-Gibbs distribution: for all i,*
$$\mu_\infty(\mathcal{U}_i) \propto \exp\left(-\tfrac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right)$$
3. *Avoidance of non-minimizers: if K_i is not minimizing, then there is K_j minimizing with $E_j < E_i$:*
$$\frac{\mu_\infty(\mathcal{U}_i)}{\mu_\infty(\mathcal{U}_j)} \le e^{-\frac{c_{i,j}}{\eta}} \qquad \textit{for some } c_{i,j} > 0$$
4. *Concentration on ground states: given $\mathcal{U}_0$ neighborhood of the ground states $K_0 = \operatorname{argmin}_i E_i$,*
$$\mu_\infty(\mathcal{U}_0) \ge 1 - e^{-\frac{c_0}{\eta}}, \qquad \textit{for some } c_0 > 0$$

Example (Gaussian noise): $E_i = \frac{f(x_i)}{2\sigma^2}$ for any $x_i \in K_i$