The Global Convergence Time of SGD in Non-Convex Landscapes
Sharp Estimates via Large Deviations

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

arXiv:2503.16398

TLDR: We characterize the average time for SGD to reach the global minimum of a non-convex function throug a large deviations 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*: $\|\nabla f(x) - \nabla f(x')\| \leq \beta \|x - x'\|$ for all x, x'
- f is *coercive*: $\lim_{\|x\| \to \infty} f(x) = +\infty$

Regularity assumption:

$\operatorname{crit}(f) := \{x \in \mathbb{R}^d \mid \nabla f(x) = 0\} = \bigcup_{i=0}^{N-1} K_i$, where K_i (smoothly) connected components

Global convergence time of SGD

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

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

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

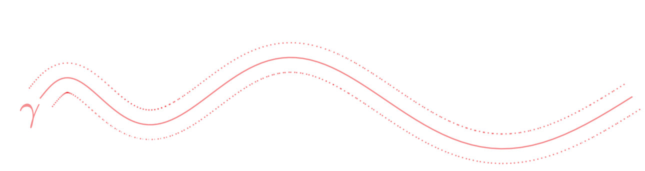
Noise assumptions:

- $\mathbb{E}[Z(x;\omega)] = 0$, $\operatorname{cov}(Z(x;\omega)) \succ 0$, $Z(x;\omega) = O(\|x\|)$
- $Z(x;\omega)$ is σ sub-Gaussian:
$$\log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right] \leq \frac{\sigma^2}{2}\|v\|^2$$
- Sufficient SNR:
$$\liminf_{\|x\| \to \infty} \frac{\|\nabla f(x)\|^2}{\sigma^2} \geq \text{some constant}$$

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; $Z(x;\omega) = \nabla f_\omega(x) - \nabla f(x)$

Large deviations for SGD

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



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

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

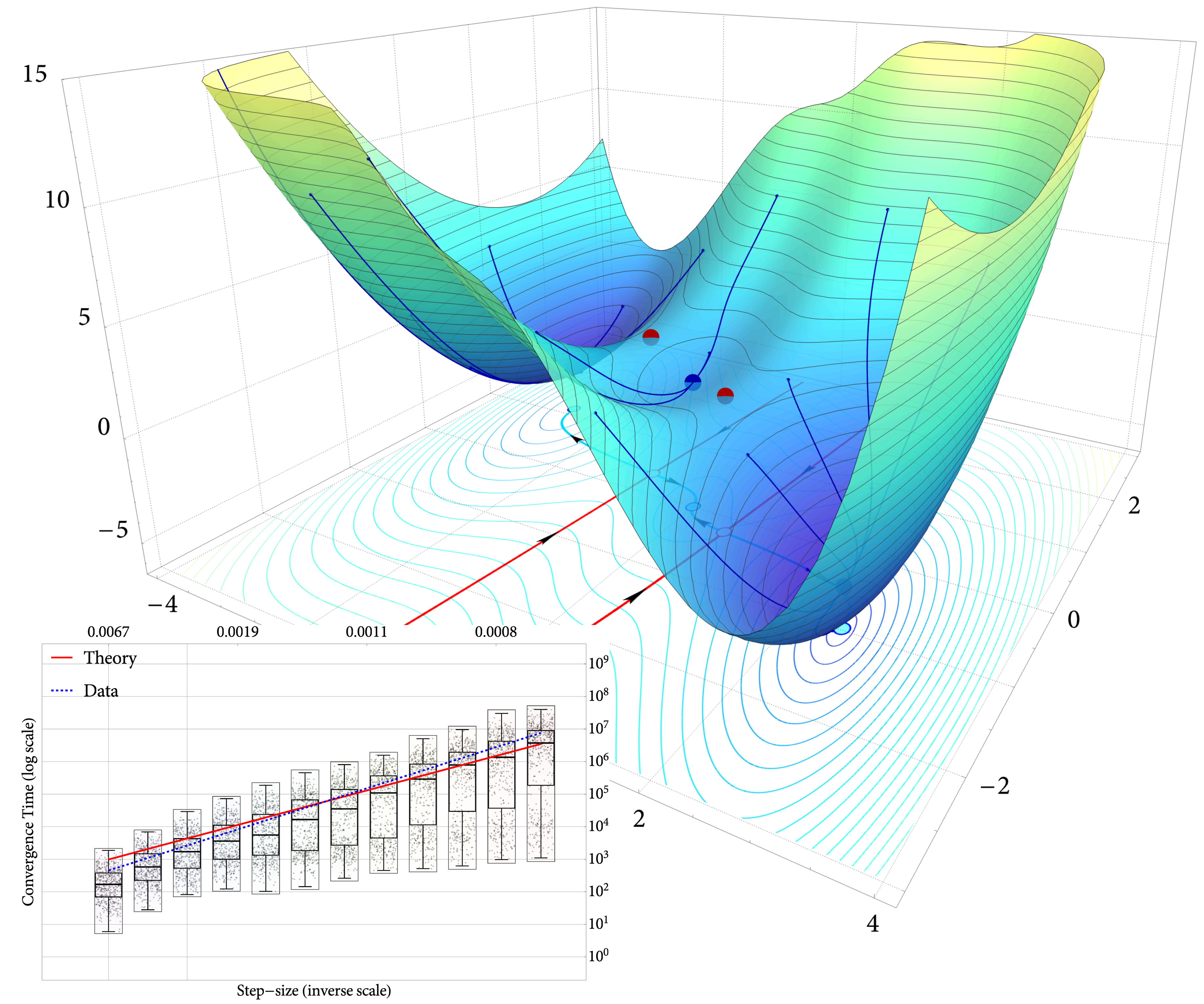
From (Azizian et al., 2024) using tools from (Freidlin & Wentzell, 1998; Dupuis, 1988)

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

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

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

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



Key findings:

- **Global convergence time of SGD:** starting at x, time τ to reach $\operatorname{argmin} f$ satisfies
$$\mathbb{E}_x[\tau] \approx \exp\left(\frac{E(x)}{\eta}\right)$$
where $E(x)$ energy of SGD starting at x
- **Key quantity $E(x)$:** geometric measure of problem's hardness, it captures
  - The difficulty of the loss landscape: hardest set of obstacles to overcome to reach $\operatorname{argmin} f$
  - The statistics of the noise: scales with inverse square of the noise level
- **Transfer of geometrical properties:** shallow local minima $\Rightarrow$ small $E(x)$

Challenges and techniques:

- Requires tools to analyze the long-run distribution of SGD in non-convex problems
- We leverage large deviation theory and the theory of random dynamical systems,
  → Estimate the probability of rare events, such as SGD switching from one local minima to another
- We adapt & refine Freidlin & Wentzell (1998); Kifer (1988), building on Azizian et al. (2024)

References

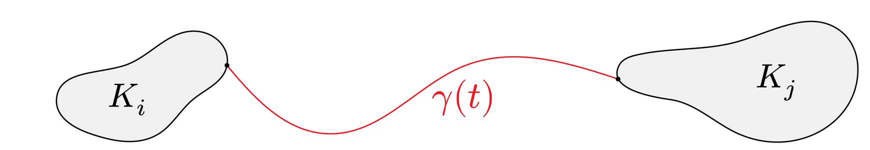
Freidlin, M. I., & Wentzell, A. D., 1998. *Random perturbations of dynamical systems.* Springer

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

Azizian, W., Iutzeler, F., Malick, J., and Mertikopoulos, P., 2024. *What is the Long-Run Distribution of Stochastic Gradient Descent? A Large Deviations Analysis.* ICML 2024

Transition between critical points

Given K_i, K_j critical points, when and how fast does SGD transition from K_i to K_j *without hitting* $\operatorname{argmin} f$?



Transition cost from K_i to K_j:

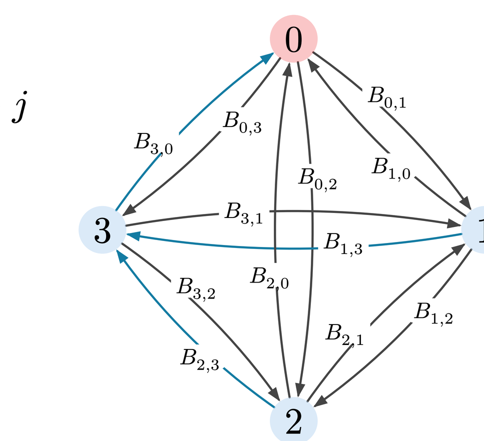
$$B_{i,j} = \inf\{\mathcal{S}_T[\gamma] \mid \gamma(0) \in K_i, \gamma(T) \in K_j, T \in \mathbb{N}, \gamma(n) \notin \operatorname{argmin} f \text{ for } n = 0, \dots, T-1\}$$

Proposition: *Transition probability from K_i to K_j without hitting $\operatorname{argmin} f$: for $\eta > 0$ small enough,*

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

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

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



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

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

Energy of pruning K_i:

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

Energy of K_0 relative to K_i:

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

Energy of K_0 relative to x:

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

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

Theorem

For any $\varepsilon > 0$, if $\eta, \delta > 0$ are small enough, then, for SGD started at x,

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

where the LHS holds under a technical condition involving the "strength of attrcation" of the $\operatorname{argmin} f$

Interpretation:

$$E(x) = 0\ \forall x \iff E(i) = 0\ \forall i \iff \text{no spurious local minima}$$

Example: Three Humps, Gaussian noise

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

Transition cost of neighboring critical points $i \to j$:

$$B_{i,j} = \frac{2\left[f(x_j) - f(x_i)\right]_+}{\sigma^2}$$

Energy of $x_0 = \operatorname{argmin} f$ relative to x near x_2:

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