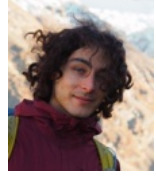
How long until SGD finds the global minimum of a non-convex function ?

Jérôme MALICK

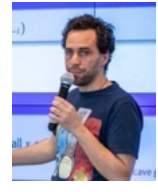
Based on joint work with

Waïss Azizian



Franck Iutzeler

Panayotis Mertikopoulos



SMAI-MODE, Nice

March 2026

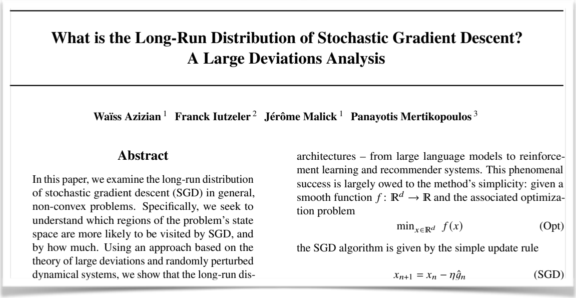
Great results, a lot of work

Waiss' 2024 ICML paper

[Azizian *et al* '24]

great, original results...

...based on a 45+ appendix !


What is the Long-Run Distribution of Stochastic Gradient Descent? A Large Deviations Analysis

Waïss Azizian [1] Franck Iutzeler [2] Jérôme Malick [1] Panayotis Mertikopoulos [3]

Abstract

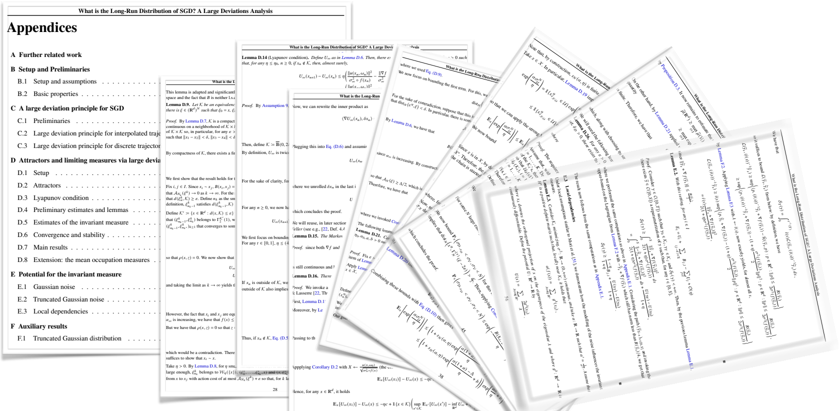
In this paper, we examine the long-run distribution of stochastic gradient descent (SGD) in general, non-convex problems. Specifically, we seek to understand which regions of the problem's state space are more likely to be visited by SGD, and by how much. Using an approach based on the theory of large deviations and randomly perturbed dynamical systems, we show that the long-run dis-

architectures – from large language models to reinforcement learning and recommender systems. This phenomenal success is largely owed to the method's simplicity: given a smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ and the associated optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \qquad \text{(Opt)}$$

the SGD algorithm is given by the simple update rule

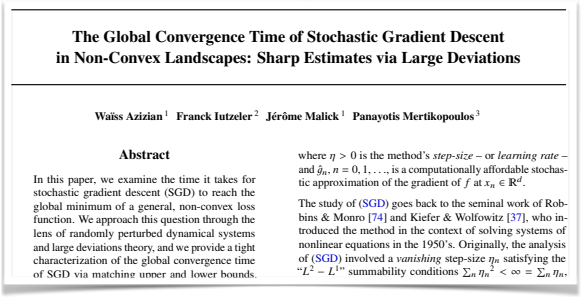
$$x_{n+1} = x_n - \eta \hat{g}_n \qquad \text{(SGD)}$$


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

Appendices

In this (follow-up) talk

Global convergence time



The Global Convergence Time of Stochastic Gradient Descent in Non-Convex Landscapes: Sharp Estimates via Large Deviations

Waïss Azizian [1] Franck Iutzeler [2] Jérôme Malick [1] Panayotis Mertikopoulos [3]

Abstract

In this paper, we examine the time it takes for stochastic gradient descent (SGD) to reach the global minimum of a general, non-convex loss function. We approach this question through the lens of randomly perturbed dynamical systems and large deviations theory, and we provide a tight characterization of the global convergence time of SGD via matching upper and lower bounds.

where $\eta > 0$ is the method's *step-size* – or *learning rate* – and $\hat{g}_n$, $n = 0, 1, \ldots$, is a computationally affordable stochastic approximation of the gradient of f at $x_n \in \mathbb{R}^d$.

The study of (SGD) goes back to the seminal work of Robbins & Monro [74] and Kiefer & Wolfowitz [37], who introduced the method in the context of solving systems of nonlinear equations in the 1950's. Originally, the analysis of (SGD) involved a *vanishing* step-size η_n satisfying the "$L^2 - L^{1+}$" summability conditions $\sum_n \eta_n^2 < \infty = \sum_n \eta_n$.

- SGD with constant stepsize:

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

- Same situation as in Waiss' talk:

  natural assumptions for smooth nonconvex optimization (⊥ Guillaume's talk yesterday)

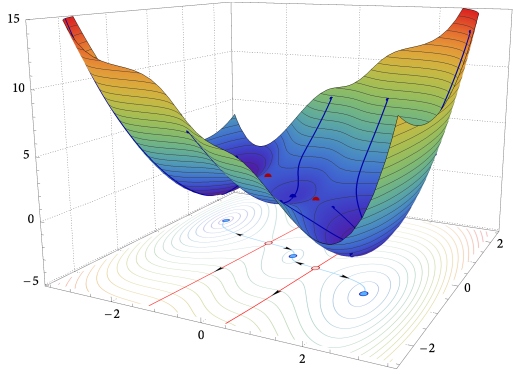
- How long for SGD to reach a global minimizer of f ?

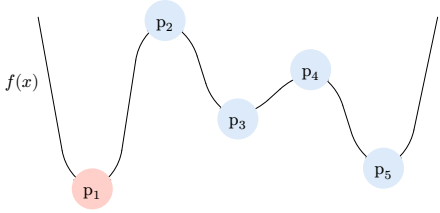
- Hitting time (with a small margin δ)

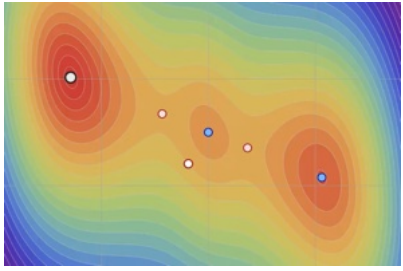
$$\tau = \min\{\ t \in \mathbb{N} :\ \mathrm{d}(\ \operatorname{argmin} f, x_t\ ) \leqslant \delta\ \}$$

- How this time depends on the stepsize η, the loss landscape, and the noise ?

Running example ("3 humps" function)







Run SGD (with Gaussian noise)

Initialized near p_3

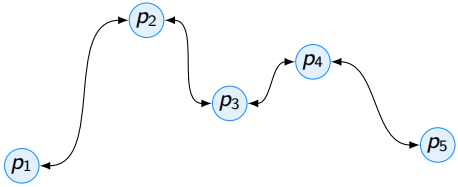
Plot one trajectory

Results applied to the example

From [Azizian *et al* '24]:

Take (simplified) transition graph

(here: $\mathcal{K}_i = \{p_i\}$)



Theorem ([Azizian *et al* '25] applied to the example)

For SGD with gaussian noise $(Z(x,\omega) \sim \mathcal{N}(0, \sigma^2 \,\mathrm{Id}))$

initialized at x *near* p_3,

$$\mathbb{E}_x[\tau] \;\approx\; \exp\left(\frac{2\big(f(p_2) - f(p_5)\big)}{\eta\,\sigma^2}\right)$$

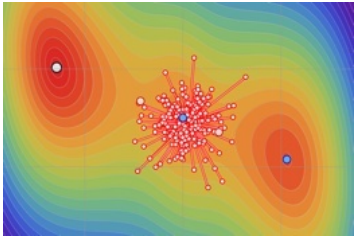


Illustration: Run 100 SGD initialized near p_3

For each trajectory: keep only last 5 iterates

stop when hit the min.

Assumptions

Assumptions for the objective

- Smoothness: $\|\nabla f(x') - \nabla f(x)\| \leqslant L\|x' - x\|$
- Coercivity: $f(x) \to \infty$ as $\|x\| \to \infty$ (+ a bit more, like norm-coercive gradient: $\|\nabla f(x)\| \to \infty$)
- Critical set regularity: $\mathrm{crit}(f)$ is a union of connected components $\mathcal{K}_i$

Assumptions for the oracle

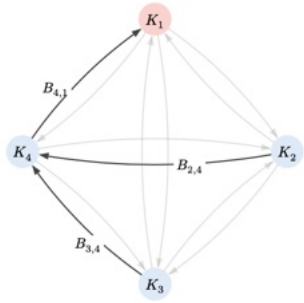
- Properness: $\mathbb{E}[Z(x;\omega)] = 0$ and $\mathrm{cov}(Z(x;\omega)) \succ 0$
- Smooth growth: Z is C^2-smooth and grows as $\|Z(x;\omega)\| \leqslant C(1 + \|x\|)$
- Light tails (Sub-Gaussian $\mathbb{P}(\|Z(x;\omega)\| > z) \leqslant \exp(-a\,z^2 + b)$)

Example: Regularized ERM: $f(x) = \frac{1}{m}\sum_{i=1}^{m} \ell(x;\xi_i) + \frac{\lambda}{2}\|x\|^2$ (with ℓ smooth)

Sample uniformly one data-point: $Z(x;\omega) = \nabla\ell(x;\xi_\omega) - \frac{1}{m}\sum_{i=1}^{m} \nabla\ell(x;\xi_i)$

Estimating the hitting time

Use the ingredients of [Azizian *et al* '24] + [Azizian *et al* '25]



- Transition graph on the $\mathcal{K}_i$'s with weights $B_{i,j}$
- "Essentially", SGD as a Markov chain on this graph
- Estimate the transitions

$$\mathbb{P}(\text{SGD goes from } \mathcal{K}_i \text{ to } \mathcal{K}_j) = \exp\left(-\frac{B_{i,j}+O(\varepsilon)}{\eta}\right)$$

$$+ \text{ average transition time} = \exp\left(\frac{\varepsilon}{\eta}\right)$$

- Energy of argmin f:

$$E_{\min} = \min\left\{\sum_{j\to k\in T} B_{j,k} \;\middle|\; T \text{ spanning tree pointing to } \mathcal{K}_1 = \operatorname{argmin} f\right\}$$

$$= \text{"the minimal cost of all the ways to the min."}$$

\+ Cost of pruning $\mathcal{K}_i$:

$$R(i \nrightarrow \min) = \min\left\{\sum_{j\to k\in T} B_{j,k} \;\middle|\; T \text{ with an edge on } i \to 1 \text{ removed}\right\}$$

$$= \text{"the minimal cost of all the ways to the min., without passing by } \mathcal{K}_i\text{"}$$

Global convergence time

Energy function $J(x)$ capturing the difficulty of the problem

- Energy of argmin f from $\mathcal{K}_i$: $J(i) = E_{\min} - R(i \nrightarrow \min)$
  $= $ "the relative difficulty to reach argmin f from $\mathcal{K}_i$"
- Energy of argmin f from x: $J(x) = \max_{i=1,\dots,N-1} [J(i) - B(x,i)]_+$
  $= $ "cost of the hardest obstacles to reach argmin f"

Theorem (Azizian *et al* '25)

Starting at x, the time τ for SGD to reach argmin f *satisfies*

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

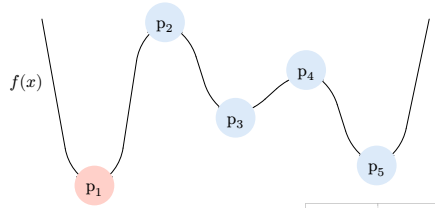
for any $\varepsilon > 0$, and $\delta, \eta > 0$ small enough.

In case of Gaussian noise

Upper-bound when $Z(x,\omega) \sim \mathcal{N}(0, \sigma^2\,\mathrm{Id})$

$$J(x) \;\leqslant\; \frac{2}{\sigma^2}\,(\text{max. extrema - min. spurious minimizer})$$

Example:



$$J(x) \;\leqslant\; \frac{2}{\sigma^2}(f(p_2) - f(p_5))$$

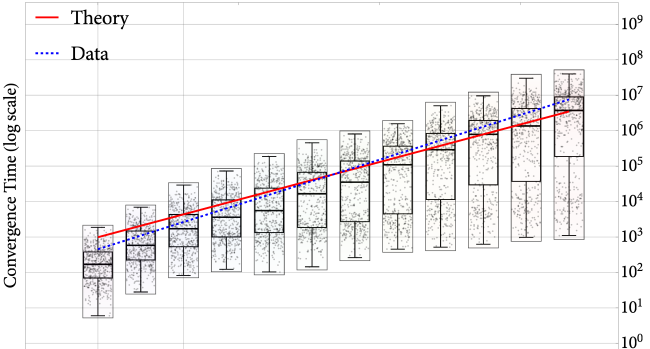
$$\log \mathbb{E}_x[\tau] \;\leqslant\; \frac{1}{\eta}\left[\frac{2}{\sigma^2}(f(p_2) - f(p_5))\right]$$

Numerical illusration:

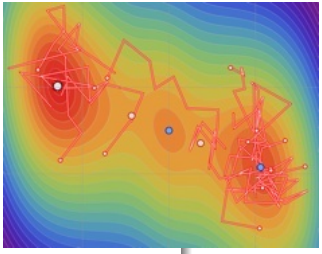
Take 12 stepsizes η

Run 10.000 SGD for each

Report hitting time (in log)



Let's sum up and look forward



Main-take aways

- SGD: famous but mysterious – now less thanks to Waiss
- Global convergence time: $\mathbb{E}_x[\tau] \approx \exp\left(\frac{J(x)}{\eta}\right)$
- Key quantity $J(x)$ measures of problem's hardness
  captures the difficulty of loss landscape and the statistics of the noise

Perspectives opened by Waiss' new tools

- Beyond SGD ? (inertia, Adam...)
- Beyond min. ? (min-max, equilibrium...)
  [Azizian, Cauvin, *et al* '26]
- Many more details and developments, here:

thank you all ☺