The long-run behaviour of SGD on nonconvex functions

Jérôme MALICK

CNRS, Grenoble, France

SIAM Optimization

Edinburgh

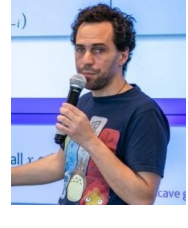
June 2026

Based on joint work with

Waïss AZIZIAN

Franck IUTZELER

Panayotis MERTIKOPOULOS

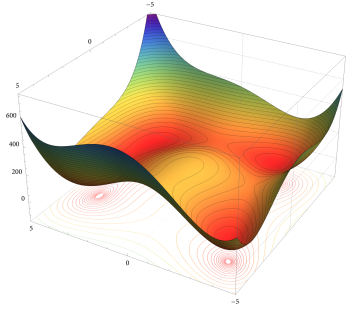


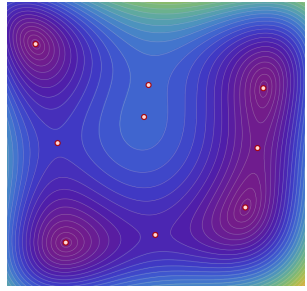
SGD does not converge...

Stochastic Gradient Descent with constant step-size $\eta > 0$

$$x_{t+1} = x_t - \underbrace{\eta}_{\text{step-size}} \widehat{g}_t \qquad \text{with} \quad \widehat{g}_t = \nabla f(x_t) + \underbrace{Z(x_t; \omega_t)}_{\text{zero-mean noise}}$$

Illustration:





Himmelblau function $\quad f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$

Video: Run 25 SGD (for each trajectory: plot only last 20 iterates)

Observation: SGD ~~converges~~ concentrates to local minimizers

What is known

SGD with constant step-size: $x_{t+1} = x_t - \eta \left[\nabla f(x_t) + Z(x_t; \omega_t) \right]$

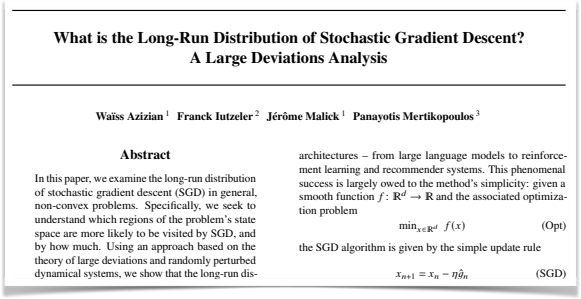
- What is known, when f is convex
  - f strongly convex: SGD converges in a ball around the minimizer [Polyak, 1987]
  - f convex: average of iterates asymptotically optimal [Polyak, Juditsky, 1992]
- What is known, when f is **non-convex**
  - In average, SGD is close to criticality [Lan, 2012] [Ghadimi, Lan, 2013]

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} ||\nabla f(x_t)||^2 \right] = O\left(\frac{1}{\sqrt{T}} \right)$$

  - Almost surely, SGD is not stuck in saddle points [Pemantle, 1990; Benaïm, Hirsch, 1995]
- What we would like to know

Question: which points are more likely to be observed — and by how much?

Our results



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 distribution of SGD resembles the Boltzmann–Gibbs distribution of equilibrium thermodynamics with temperature equal to the method's step size and energy levels determined by the problem's objective and the structure of its critical manifolds. As an application of our theory, we analyze the behavior of SGD in deep learning 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

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

Theorem (Azizian et al., 2024; very informal)

1. The **critical set** of f is visited exponentially more often than any **non-critical** set
2. Critical points are visited according to a **Boltzmann–Gibbs** distribution with temp. η
3. **Minimizers** are visited exponentially more often than **non-minimizers**
4. The **ground state** of the problem is visited exponentially more often than any other min.

Goal of this talk: make these results more precise

The way to these results

- **Key object**: asymptotic distributions of x_t
  - **Occupation measure:** average time spent by the iterates in a set of interest $\mathcal{B}$

$$\mu_t(\mathcal{B}) = \mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t} \mathbf{1}\{x_s \in \mathcal{B}\}\right]$$

  - Its limit is an **invariant measure:** $\quad x_t \sim \mu_\infty \implies x_{t+1} \sim \mu_\infty$

Question: where does μ_∞ concentrate?

The way to these results

- **Key object**: asymptotic distributions of x_t
  - **Occupation measure:** average time spent by the iterates in a set of interest $\mathcal{B}$

$$\mu_t(\mathcal{B}) = \mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t} \mathbf{1}\{x_s \in \mathcal{B}\}\right]$$

  - Its limit is an **invariant measure:** $\quad x_t \sim \mu_\infty \implies x_{t+1} \sim \mu_\infty$

Question: where does μ_∞ concentrate?

- **Approach:** Randomly perturbed dynamical systems [Freidlin Wenzel 1998, Kiffer 1980]
  $\rightarrow$ estimate the probability of rare events, such as SGD escaping a local minimum

- **Refinement:** We adapt this theory, with three main challenges:

  Lack of compactness + Noise models (finite sum) + Discrete-time dynamics

What we are not doing

- **Stochastic approximation**:

$$x_{t+1} = x_t - \eta_t \left[\nabla f(x_t) + Z(x_t; \omega_t)\right] \quad \text{with} \quad \eta_t \text{ vanishing}$$

Convergence to local minima [Bertsekas & Tsitsiklis '00, Mertikopoulos et al. '20]
but no information about which one

- **Langevin dynamics**:

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

Convergence of the distribution of the iterates to e^{-f/σ^2} [Raginsky et al. '17]
but scaling of the noise differs from SGD + we want non-gaussian noise

- **Continuous-time dynamics** (Gradient flow, SDE):

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

Results on stationary distribution + finite-time approximation of SGD [Li et al. '17]
but unsuitable for concentration questions

Existing analysis do not carry over

Blanket assumptions

Assumptions on 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 coercive gradient: $\|\nabla f(x)\| \to \infty$)
- Critical set: $\mathrm{crit}(f) = K_1 \cup \cdots \cup K_p$ with smoothly connected components K_i

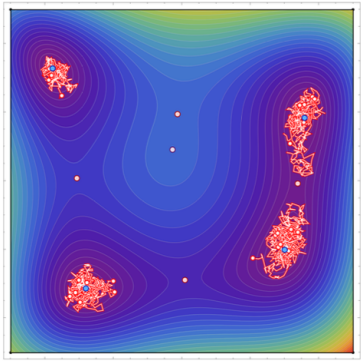
Assumptions on the noise

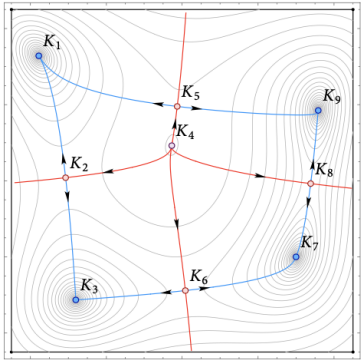
- Properness: $\mathbb{E}[Z(x;\omega)] = 0$ and $\mathrm{cov}(Z(x;\omega)) \succ 0$
- Smooth growth: Z is C^2-smooth and $\|Z(x;\omega)\| \leqslant C(1 + \|x\|)$
- Sub-Gaussian: $\log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right] \leqslant \frac{\sigma^2}{2}\|v\|^2$ ($\Rightarrow$ bounds $\mathbb{P}(\|Z(x;\omega)\| > z) \leqslant \exp(-a\,z^2 + b)$)

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

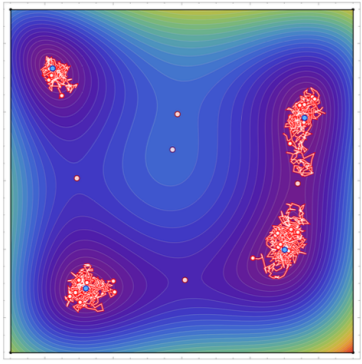
Sample one (or several) data-point ξ_ω, and undergo $Z(x;\omega) = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} \nabla\ell(x;\xi_i) - \nabla\ell(x;\xi_\omega)$

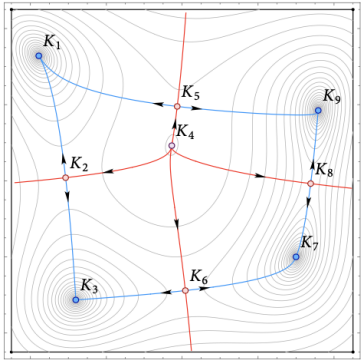
Main idea of the proof





Main idea of the proof





the long-run distribution of SGD resembles
a finite-state Markov chain on the critical components

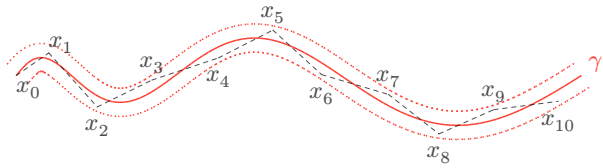
- Make this precise (difficult) – 100 pages of maths in one slide: next slide
- Get the results for SGD from the induced chain (easier) – the following slide

Key ingredient: large deviations for SGD

For $T > 0$

Continuous path $\gamma : [0, T] \to \mathbb{R}^d$

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



Proposition (Azizian et al., 2024; informal)

For small η:
$$\mathbb{P}\left(\text{SGD on } [0, T/\eta] \approx \gamma\right) \approx \exp\left(-\frac{S_T[\gamma]}{\eta}\right)$$

C. Gen. Fct: $\mathcal{H}(x, v) = \log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right]$

Conj.: $\mathcal{L}(x, v) = \mathcal{H}^*\left(x, -v - \nabla f(x)\right)$

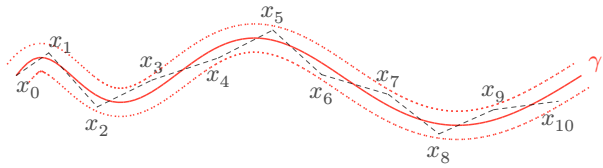
Action: $S_T[\gamma] = \int_0^T \mathcal{L}(\gamma_t, \dot{\gamma}_t)\, dt$

Key ingredient: large deviations for SGD

For $T > 0$

Continuous path $\gamma : [0, T] \to \mathbb{R}^d$

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



Proposition (Azizian et al., 2024; informal)

For small η: $$\mathbb{P}\left(\text{SGD on } [0, T/\eta] \approx \gamma\right) \approx \exp\left(-\frac{S_T[\gamma]}{\eta}\right)$$

		Gaussian noise $(Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I_d))$
C. Gen. Fct:	$\mathcal{H}(x, v) = \log \mathbb{E}\left[e^{\langle v, Z(x;\omega)\rangle}\right]$	$\mathcal{H}(x, v) = \frac{\sigma^2}{2}\|v\|^2$
Conj.:	$\mathcal{L}(x, v) = \mathcal{H}^*\left(x, -v - \nabla f(x)\right)$	$\mathcal{L}(x, v) = \frac{\|v+\nabla f(x)\|^2}{2\sigma^2}$
Action:	$S_T[\gamma] = \int_0^T \mathcal{L}(\gamma_t, \dot\gamma_t)\, dt$	$S_T[\gamma] = \frac{1}{2\sigma^2}\int_0^T \|\dot\gamma_t + \nabla f(\gamma_t)\|^2 dt$

From transitions to energy

We get (...) the transition cost from one critical region to another

$$B_{i,j} = \min\{S_T[\gamma] \ : \ \gamma(0) \in K_i, \ \gamma(T) \in K_j, \ T > 0\}$$

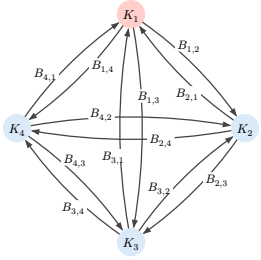
Then, using exact formulas for finite-state Markov chains, we get (...):

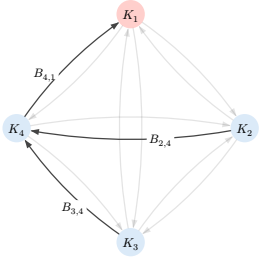
Lemma (very informal)

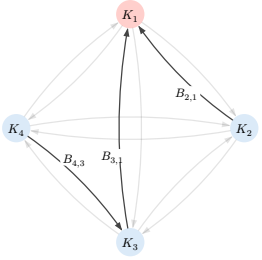
The invariant measure of SGD restricted to $\{K_1, \dots, K_p\}$ is, for $\eta > 0$ small enough

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

where energy of K_i: $\quad E_i = \min\left\{\sum_{j \to k \in T} B_{j,k} \ : \ T \text{ spanning tree pointing to } i\right\}$







From energy to invariant distribution

Theorem (Azizian et al., 2024; less informal)

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

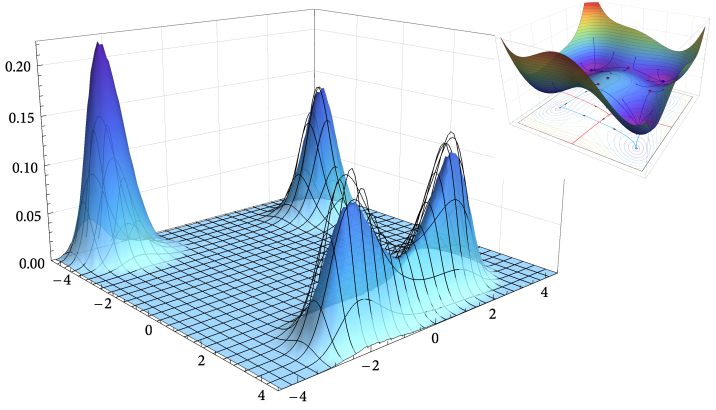
1. **Concentration near critical points:** there is some $c > 0$ s.t.
$$\mu_\infty \left(\bigcup_{i=1}^{p} \mathcal{U}_i \right) \geqslant 1 - e^{-c/\eta} \qquad \text{for some } c > 0$$
2. **Boltzmann–Gibbs distribution:** for all i,
$$\mu_\infty(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + O(\varepsilon)}{\eta} \right)$$
3. **Saddle-point avoidance:** if K_i is a saddle, then there is K_j local minimum with $E_j < E_i$
$$\frac{\mu_\infty(\mathcal{U}_i)}{\mu_\infty(\mathcal{U}_j)} \approx e^{-(E_i - E_j)/\eta}$$
4. **Ground state concentration:** $\mathcal{U}_0$ neighborhood of the ground state $K_0 = \operatorname{argmin}_i E_i$
$$\mu_\infty(\mathcal{U}_0) \geqslant 1 - e^{-c/\eta} \qquad \text{for some } c > 0$$

Illustration: Gaussian noise

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

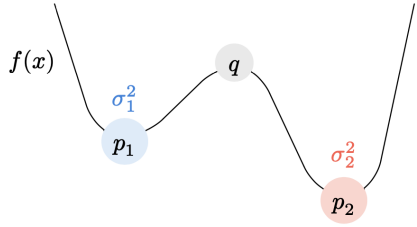
Compute $E_i = \dfrac{2f(K_i)}{\sigma^2} \implies \mu_\infty(\mathcal{U}_i) \approx \exp\left(-\dfrac{2f(K_i)}{\sigma^2 \eta}\right)$ and $K_0 = \text{global min.}$

Illustration: run 1000 SGD on Himmelblau, observe distribution of iterates at $T = 1000$



Energy landscape (wireframe) vs. empirical distribution of SGD (blue)

Remark: ground state = global minimizer ?



$$E_1 = \frac{f(q) - f(p_2)}{\sigma_2^2} \qquad \text{and} \qquad E_2 = \frac{f(q) - f(p_1)}{\sigma_1^2}$$

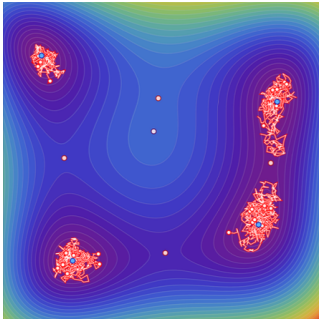
$$\sigma_1 \text{ small enough} \implies E_1 < E_2 \implies \mu_\infty(p_1) \ll \mu_\infty(p_2)$$

In general, minimizer of the energy $\neq$ minimizer of the function

Let's sum up and look forward

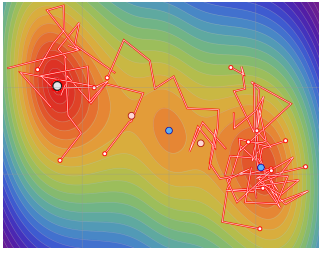
Main take-aways

- SGD concentrates near local minima
  in particular, near those that minimize the energy
- We developed a new theoretical framework
  to analyze SGD with the help of large deviations



Perspectives

- Beyond SGD ? (inertia, Adam...)
- Beyond minimization ? (min-max, equilibrium...)
- Beyond distribution ? Eg. Hitting time [Azizian et al '25]



thank you all ☺