What is the Long-Run Behavior of Stochastic Gradient Descent?

A Large Deviation Analysis

Franck IUTZELER
Université Toulouse III

Workshop “Recent developments beyond classical regimes in statistical learning”, Toulouse
November 8, 2024

- Training of deep neural networks ≈ SGD on a nonconvex loss function
- Lots of minimizers and lots of randomness (initialisation, mini-batching, etc)

Image credit: losslandscape.com

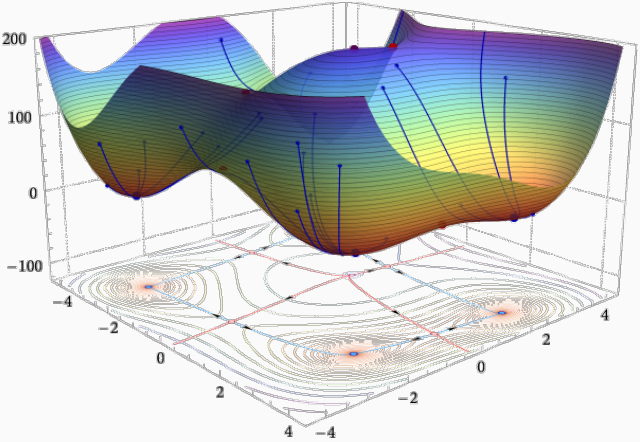
- **Objective function** $f : \mathbb{R}^d \to \mathbb{R}$ smooth nonconvex
- **Stochastic Gradient Descent (SGD)** with constant step-size

$$x_{n+1} = x_n - \underbrace{\eta}_{\text{stepsize}} \left[\nabla f(x_n) + \underbrace{Z(x_n; \omega_{n+1})}_{\text{zero-mean noise}} \right]$$

Question: What is the **asymptotic behavior** of SGD?

Running example • Himmelblau function

- $f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$



- constant stepsize + noise ⇝ no pointwise convergence

Literature • SGD w/ constant stepize

- Lines of work that **do not characterize the asymptotic behavior**
  - Stochastic Approximation when $\eta_n \propto n^{-(1+\varepsilon)}$ convergence to local minima but no information about which one [Bertsekas & Tsitsiklis, 2000]
  - Sampling (MCMC, Langevin) scaling of the noise differs from SGD
  - Continuous-time limit (SDE) only valid on finite time horizons

- **Classical results** in optimization
  - f convex average of SGD iterates is near-optimal
  - f nonconvex near-critical in average $\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\|\nabla f(x_n)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ [Lan, 2012] and avoids saddle points [Brandière & Duflo, 1996; Mertikopoulos et al., 2020]

This presentation • Long run behavior of SGD

- **Which critical points** (local minima) are visited the most in the long run?
- Theory of **large deviations** and random perturbations of dynamical systems
  - Estimate the probability of rare events, such as SGD escaping a local minima
- (Almost) **Realistic assumptions** on the noise and objective

- Joint work with Waïss Azizian, Panayotis Mertikopoulos, Jérôme Malick
  - arXiv 2406.09241 ICML 2024

Setup & Assumptions

- Objective function f
  - smooth C^2 and ∇f is β-Lipschitz continuous
  - coercive $\lim_{\|x\|\to\infty} f(x) = +\infty$
  - gradient coercive $\lim_{\|x\|\to\infty} \|\nabla f(x)\| = +\infty$

- Noise term Z
  - proper $\mathbb{E}[Z(x;\omega)] = 0$ and $\operatorname{cov}(Z(x;\omega)) \succ 0$ for all $x \in \mathbb{R}^d$
  - limited growth $Z(x;\omega) = \mathcal{O}(\|x\|)$ almost surely
  - sub-Gaussian $\log \mathbb{E}[\exp(\langle p, Z(x;\omega)\rangle)] \leq \frac{\sigma^2\|p\|^2}{2}$

Recall SGD

$$x_{n+1} = x_n - \eta\,[\nabla f(x_n) + Z(x_n;\omega_{n+1})]$$

- Objective function f
  - smooth — C^2 and ∇f is β-Lipschitz continuous
  - coercive — $\lim_{\|x\| \to \infty} f(x) = +\infty$
  - gradient coercive — $\lim_{\|x\| \to \infty} \|\nabla f(x)\| = +\infty$

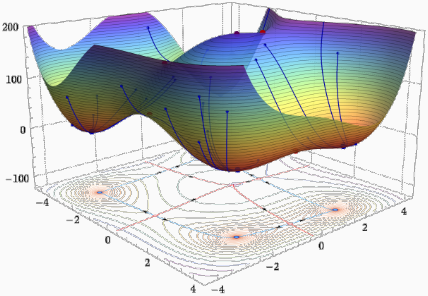
- Noise term Z
  - proper — $\mathbb{E}[\mathrm{Z}(x;\omega)] = 0$ and $\mathrm{cov}(\mathrm{Z}(x;\omega)) \succ 0$ for all $x \in \mathbb{R}^d$
  - limited growth — $\mathrm{Z}(x;\omega) = \mathcal{O}(\|x\|)$ almost surely
  - sub-Gaussian — $\log \mathbb{E}[\exp(\langle p, \mathrm{Z}(x;\omega)\rangle)] \le \frac{\sigma^2 \|p\|^2}{2}$

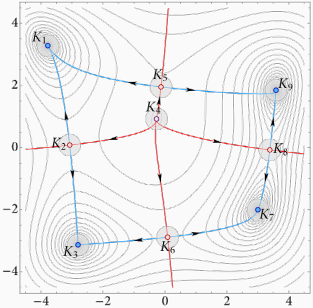
Example Regularized ERM $\quad f(x) = \frac{1}{m} \sum_{i=1}^m \ell(x;\xi_i) + \frac{\lambda}{2}\|x\|^2$

SGD by sampling one example leads to $\mathrm{Z}(x;\omega) = \nabla \ell(x;\xi_\omega) - \frac{1}{m}\sum_{i=1}^m \nabla \ell(x;\xi_i)$

where ω is sampled uniformly at random in $\{1, .., m\}$.

- Critical set $\mathrm{crit}(f) := \{x \in \mathbb{R}^d : \nabla f(x) = 0\}$
  - finite number of smoothly connected components $\mathrm{crit}(f) = \{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_K\}$





Not that restrictive Holds for definable functions

- We focus on the **invariant measure** μ_∞^η of SGD
  - defining property

$$x \sim \mu_\infty^\eta \implies x - \eta \left[\nabla f(x) + Z(x;\omega)\right] \sim \mu_\infty^\eta$$

  - weak* limit of the mean occupation measure

$$\mu_n(\mathcal{B}) = \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{1}\{x_k \in \mathcal{B}\}\right]$$

- We analyze the **relative measures** of the critical components $\{\mathcal{K}_i\}_{i=1}^K$
  - Concentration near minimizers as $\eta \to 0$
  - Comparison of critical components $\mu_\infty^\eta(\mathcal{K}_i)/\mu_\infty^\eta(\mathcal{K}_j)$

Discrete $\leftrightarrow$ Continuous Time & Large Deviations Approach

$$x_{n+1} = x_n - \eta\left[\nabla f(x_n) + Z(x_n;\omega_{n+1})\right] = x_0 - \eta\sum_{k=0}^{n} \nabla f(x_k) + Z(x_k;\omega_k)$$

- Markov chain
  - (weak) Feller $\Rightarrow$ existence of an invariant measure [Douc et al., 2018]
  - No useful characterization of the invariant measure known
- **"Discrete-time" Large deviation principle** by Cramér's theorem

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=0}^{n}\nabla f(x) + Z(x;\omega_k) \in \mathcal{B}\right] \sim_{n\to\infty} \exp\left(-n \inf_{v\in\mathcal{B}} \mathcal{L}(x,v)\right)$$

  - Characterizes the probability of staying in any Borel $\mathcal{B}$ and in particular minimizers neighborhoods...
  - Relies on some Lagrangian function (more later)

$$x_{n+1} = x_n - \eta\,[\nabla f(x_n) + Z(x_n;\omega_{n+1})] = x_0 - \eta \sum_{k=0}^{n} \nabla f(x_k) + Z(x_k;\omega_k)$$

- Markov chain
  - (weak) Feller $\Rightarrow$ existence of an invariant measure [Douc et al., 2018]
  - No useful characterization of the invariant measure known
- **"Discrete-time" Large deviation principle** by Cramér's theorem

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=0}^{n} \nabla f(x) + Z(x;\omega_k) \in \mathcal{B}\right] \sim_{n\to\infty} \exp\left(-n \inf_{v\in\mathcal{B}} \mathcal{L}(x, v)\right)$$

  - Characterizes the probability of staying in any Borel $\mathcal{B}$ and in particular minimizers neighborhoods... But in SGD, x is not fixed but highly correlated!
  - Relies on some Lagrangian function (more later)

- Discrete time

$$x_{n+1} = x_n - \eta \left[\nabla f(x_n) + Z(x_n; \omega_{n+1})\right]$$

- Continuous time
  - "interpolated" trajectory for any $n \geq 0$, $t \in [\eta n, \eta(n+1)]$

$$X_t = x_n + \left(\frac{t}{\eta} - n\right)(x_{n+1} - x_n)$$

  - continuous "discretized noise" trajectory for any $t > 0$ with $Z_0 = x_0$

$$\dot{Z}_t = -\nabla f(Z_t) + Z(Z_t, \omega_{\lfloor t/\eta \rfloor})$$

Remarks X_t is natural but Z_t goes better with Lagrangians in the analysis

Time is accelerated as $\Delta t = 1 \leftrightarrow \Delta n = 1/\eta$ to have "enough noise" from t to $t+1$

The SDE $\dot{Y}_t = -\nabla f(Y_t) + U(Y_t)\, dW_t$ is different, has the wrong scale for the noise ($\sqrt{\eta}$ instead of η), and the discretization or the convergence is exponentially bad in η [Raginsky et al., 2017 ; Li et al., 2019]

- Idea inspired from [Freidlin and Wentzell, 1998]
  - $\{0, 1/\eta, .., T/\eta\}$ iterates of SGD $\approx [0, T]$ trajectory of $\dot{Z}_t = -\nabla f(Z_t) + Z(Z_t, \omega_{\lfloor t/\eta \rfloor})$
  - Trajectory of Z_t is a point in the space of continuous curves $C_T := C([0, T], \mathbb{R}^d)$
  - Derive a large deviations principle for curves $\gamma \in C_T$
- Ingredients
  - Cumulant Generating Function $K(x, p) := \log \mathbb{E}[\exp(\langle p, Z(x; \omega)\rangle)] + \langle \nabla f(x), p\rangle$
  - Lagrangian $\mathcal{L}(x, v) := K^*(x, -v)$ is its convex conjugate (in v)
  - Action functional $\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}(\gamma(t), \dot{\gamma}(t))\, dt$

Proposition As $\eta \to 0$,

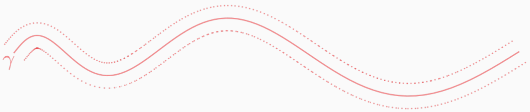
$$\text{As } \eta \to 0, \quad \mathbb{P}\left(\frac{T}{\eta} \text{ steps of SGD} \approx \gamma\right) \approx \mathbb{P}\left(\mathrm{dist}_{0,T}(Z, \gamma) < \delta\right) \approx \exp\left(-\frac{\mathcal{S}_T[\gamma]}{\eta}\right)$$

Gaussian case $\mathcal{L}(x, v) = \frac{\|v + \nabla f(x)\|^2}{2\sigma^2}$ and $\mathcal{S}_T[\gamma] = \int_0^T \frac{\|\dot{\gamma}(t) + \nabla f(\gamma(t))\|^2}{2\sigma^2}\, dt$

Proposition As $\eta \to 0$

$$\mathbb{P}\left(\frac{T}{\eta} \text{ steps of SGD} \approx \gamma\right) \approx \exp\left(-\frac{\mathcal{S}_T[\gamma]}{\eta}\right)$$

- Interpretation
  - Trajectories of SGD tend to concentrate near action-minimizing curves
  - Gradient flows are privileged as $\mathcal{L}(x, v) \geq 0$ and $\mathcal{L}(x, v) = 0 \iff v = \nabla f(x)$



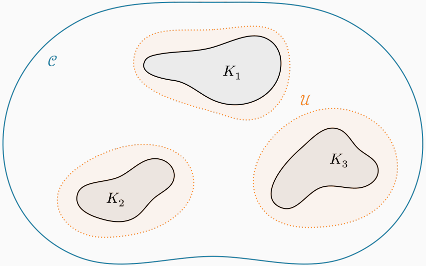
Proposition As $\eta \to 0$

$$\mathbb{P}\left(\frac{T}{\eta} \text{ steps of SGD} \approx \gamma\right) \approx \exp\left(-\frac{\mathcal{S}_T[\gamma]}{\eta}\right)$$

- **What about critical components?** $\operatorname{crit}(f) := \{x \in \mathbb{R}^d : \nabla f(x) = 0\} = \{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_K\}$
  - SGD does concentrates on critical points by following the flow
  - Next step is to compare paths between critical components

Lemma Given $\operatorname{crit}(f) \subset \mathcal{U} \subset \mathcal{C}$ with $\mathcal{U}$ open, $\mathcal{C}$ compact, for $\eta > 0$ small enough

$$\mathbb{P}(\text{ SGD reaches } \mathcal{U} \text{ in} \geq n \text{ steps}) \leq e^{-\Omega(n/\eta)}$$



Transitions between critical components

- **Definition** following [Kifer, 1988]

$$B(x,x') \coloneqq \inf\{\mathcal{S}_T[\gamma] : \gamma \in \mathcal{C}_T, \gamma(0) = x, \gamma(T) = x', T \in \mathbb{N}\}$$

  - **fixes** some transition **time** T
  - if there is a **gradient flow** going from x to x', then $B(x,x') = 0$
  - **equivalence classes** of $x \sim x' \iff B(x,x') = B(x',x) = 0$ are $\{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_K\}$

- Potentials for transitioning **between critical components**

$$B_{ij} \coloneqq \inf\{\mathcal{S}_T[\gamma] : \gamma \in \mathcal{C}_T, \gamma(0) \in \mathcal{K}_i, \gamma(T) \in \mathcal{K}_j, T \in \mathbb{N}\}$$

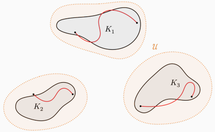
  - From **Step 1**, we have for $\eta > 0$ small enough

$$\mathbb{P}(\text{SGD transitions from } \mathcal{K}_i \text{ to } \mathcal{K}_j) \approx \exp\left(-\frac{B_{ij}}{\eta}\right)$$

- Consider the **homogeneous discrete chain** on $\{1, .., K\}$

$$z_n = i \text{ if the } n\text{-th visited component is } \mathcal{K}_i \text{ (up to a small neighborhood)}$$

  - From Step 1, critical neighborhoods are exponentially more visited so the **invariant distribution of z_n** captures the long-run behavior of SGD
  - **Transitions probabilities** are given by the B_{ij}



Lemma The invariant distribution π of z_n for $\eta > 0$ small enough satisfies

$$\pi(i) \propto \exp\left(-\frac{E_i}{\eta}\right) \quad \text{with} \quad E_i = \min_{T_i \in \mathcal{T}_i} \sum_{j,k \in T_i} B_{jk}$$

the **energy** of $\mathcal{K}_i$ defined as the minimal weight of a spanning tree rooted at i

Main Result

Theorem Given $\varepsilon > 0$ and $\mathcal{U}_i$ sufficiently small neighborhoods of the components of $\operatorname{crit}(f)$. Then, for sufficiently small $\eta > 0$, we have

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

$$\mu_\infty^\eta(\cup_{i=1}^K \mathcal{U}_i) \geq 1 - e^{-\lambda/\eta}$$

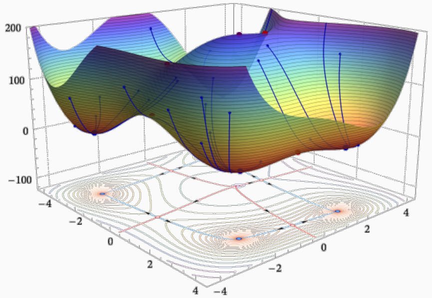
- **Boltzmann-Gibbs distribution** for all i

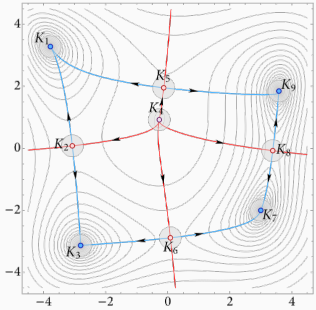
$$\mu_\infty^\eta(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right)$$

- **Concentration on ground states** given $\mathcal{U}_0$ neighborhood of $\arg\min_i E_i$

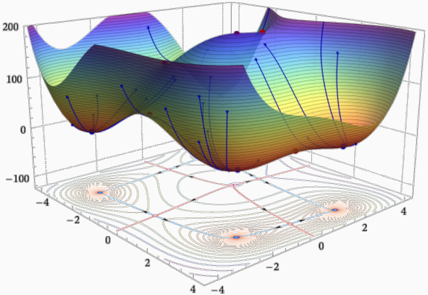
$$\mu_\infty^\eta(\mathcal{U}_0) \geq 1 - e^{-\lambda_0/\eta} \quad \text{for some} \quad \lambda_0 > 0$$

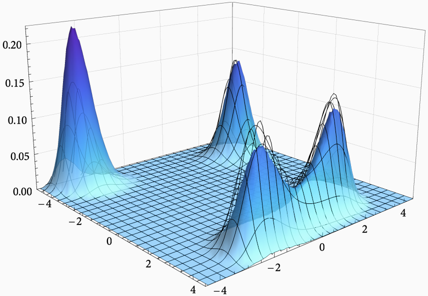
- Assume that $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I)$
  - $E_i = 2f(x_i)/\sigma^2$ for any $x_i \in \mathcal{K}_i$
  - $B_{51} = 0 \quad B_{15} = 2(f(x_5) - f(x_1))/\sigma^2$ for $(x_1, x_5) \in \mathcal{K}_1 \times \mathcal{K}_5$



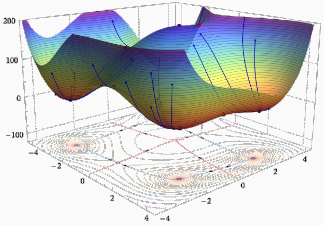


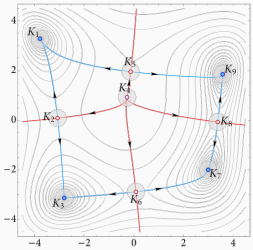
- Assume that $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I)$
  - $E_i = 2f(x_i)/\sigma^2$ for any $x_i \in \mathcal{K}_i$
  - $B_{51} = 0 \quad B_{15} = 2(f(x_5) - f(x_1))/\sigma^2$ for $(x_1, x_5) \in \mathcal{K}_1 \times \mathcal{K}_5$

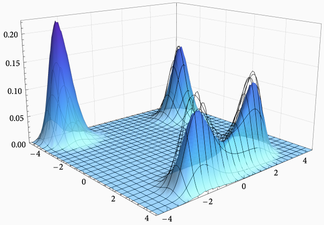




- We introduce a theory of **large deviations** for SGD in nonconvex problems
    - Sound approach for the long-run of SGD
    - Precise adaptation of random perturbations of dynamical systems' theory
- We characterize the **asymptotic distribution of SGD**
    - Critical regions are visited exponentially more often than non-critical regions
    - Critical components are visited with probability exponentially proportional to their energy, not necessarily their function value
- **Future steps** in the comprehension of stochastic methods in nonconvex landscapes
    - More realistic algorithms (momentum, adam)
    - Links with neural networks landscape and generalization







Thank you for your attention