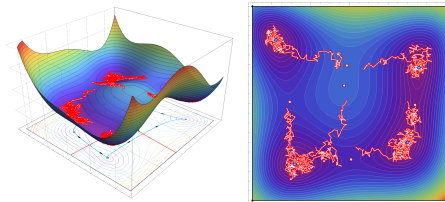


THE LONG-RUN BEHAVIOR OF SGD IN NON-CONVEX LANDSCAPES

LARGE DEVIATIONS, SOLUTION SELECTION, AND MORE



Panayotis Mertikopoulos

French National Center for Scientific Research (CNRS)

⟨ Inria @ UniCA | Selected Lectures on Statistical ML | February 25, 2026 ⟩

About

- 📄 M, Hallak, Kavis & Cevher, *On the almost sure convergence of stochastic gradient descent in non-convex problems*, NeurIPS 2020
- 📄 Azizian, Iutzeler, Malick & M, *What is the long-run distribution of stochastic gradient descent? A large deviations analysis*, ICML 2024
- 📄 Azizian, Iutzeler, Malick & M, *The global convergence time of stochastic gradient descent in non-convex landscapes: Sharp bounds via large deviations*, ICML 2025



N. Hallak



A. Kavis



V. Cevher



arxiv/2006.11144

About

- 📄 M, Hallak, Kavis & Cevher, *On the almost sure convergence of stochastic gradient descent in non-convex problems*, NeurIPS 2020
- 📄 Azizian, Iutzeler, Malick & M, *What is the long-run distribution of stochastic gradient descent? A large deviations analysis*, ICML 2024
- 📄 Azizian, Iutzeler, Malick & M, *The global convergence time of stochastic gradient descent in non-convex landscapes: Sharp bounds via large deviations*, ICML 2025



W. Azizian



F. Iutzeler



J. Malick



arxiv/2406.09241



arxiv/2503.16398

Outline

① Background & prelims

② Part I: SGD with a vanishing step-size

③ Part II: SGD with a constant step-size

Deep nets

Training a neural net: $\min_{W \in \mathbb{R}^n} \text{ERM}(W) = \frac{1}{|\text{Data}|} \sum_{X \in \text{Data}} \|\text{Truth}(X) - \text{Out}(W; X)\|^2$

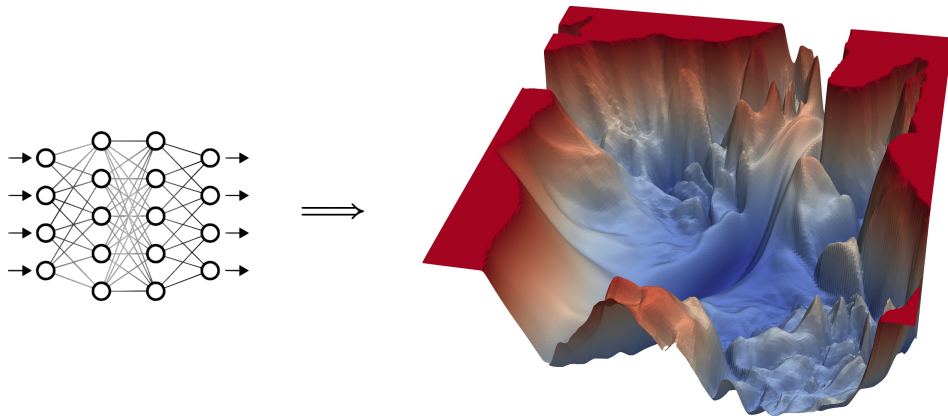
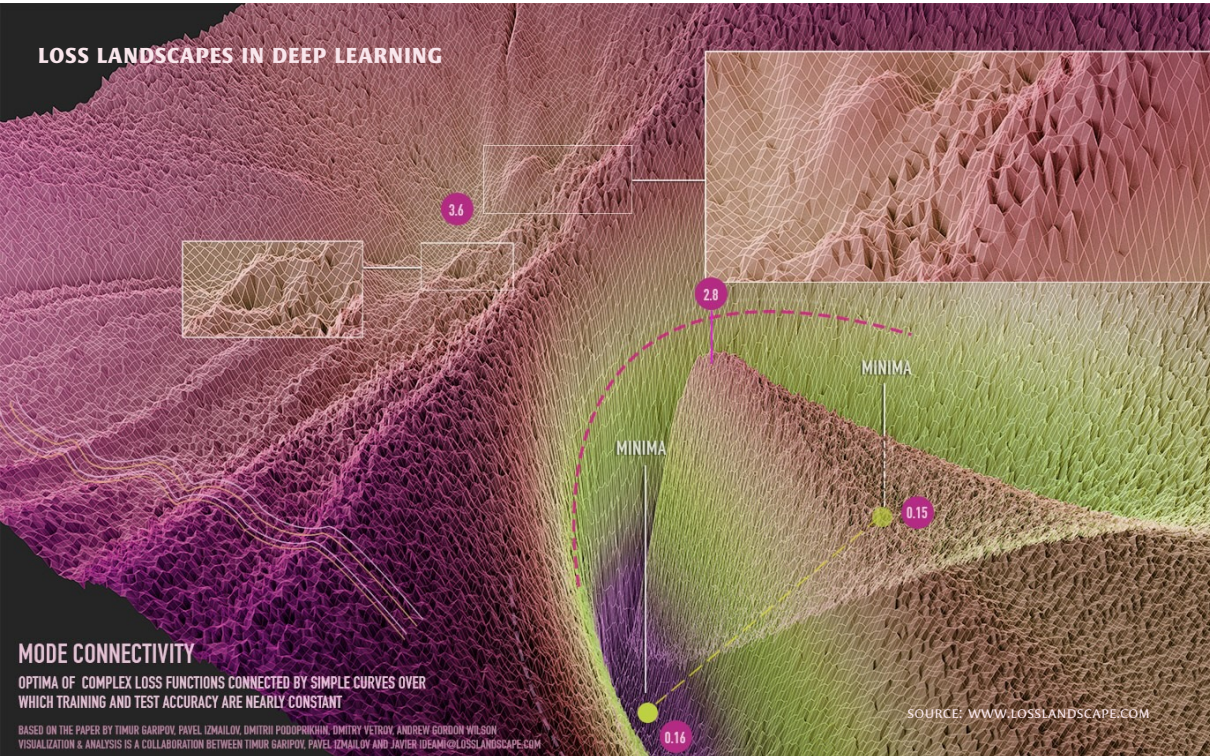


Figure: The loss landscape of a deep neural network (Li et al., 2018)

LOSS LANDSCAPES IN DEEP LEARNING



MODE CONNECTIVITY

OPTIMA OF COMPLEX LOSS FUNCTIONS CONNECTED BY SIMPLE CURVES OVER WHICH TRAINING AND TEST ACCURACY ARE NEARLY CONSTANT

BASED ON THE PAPER BY TIMUR GARIPOV, PAVEL IZMAILOV, DMITRII PODOPRIKHIN, DMITRY VETROY, ANDREW GORDON WILSON
VISUALIZATION & ANALYSIS IS A COLLABORATION BETWEEN TIMUR GARIPOV, PAVEL IZMAILOV AND JAVIER IDEAMI@LOSSLANDSCAPE.COM

SOURCE: WWW.LOSSLANDSCAPE.COM

LOSS LANDSCAPES IN DEEP LEARNING



SOURCE: WWW.LOSSLANDSCAPE.COM

LOSS LANDSCAPES IN DEEP LEARNING

A 3D visualization of a loss landscape, showing a rugged, mountainous terrain with a bright, glowing pink and yellow light source in the center, representing the minimum of the loss function. The landscape is composed of dark, rocky peaks and valleys, with the central light source illuminating the surrounding area.

SOURCE: WWW.LOSSLANDSCAPE.COM

Stochastic optimization

Stochastic minimization problems

$$\underset{x \in \mathcal{X}}{\text{minimize}} \ f(x) = \mathbb{E}_{\theta}[F(x; \theta)] \quad (\text{Opt})$$

e.g., finite sum: $f(x) = \sum_{i=1}^N f_i(x; \theta_i)$

Stochastic optimization

Stochastic minimization problems

$$\underset{x \in \mathcal{X}}{\text{minimize}} \ f(x) = \mathbb{E}_{\theta}[F(x; \theta)] \quad (\text{Opt})$$

e.g., finite sum: $f(x) = \sum_{i=1}^N f_i(x; \theta_i)$

Stochastic gradients

Access to f provided by *stochastic gradients* of the form

$$G(x; \omega) = \nabla f(x) + Z(x; \omega) \quad (\text{SG})$$

minibatch sampling, ...

Stochastic optimization

Stochastic minimization problems

$$\underset{x \in \mathcal{X}}{\text{minimize}} \ f(x) = \mathbb{E}_{\theta}[F(x; \theta)] \quad (\text{Opt})$$

e.g., finite sum: $f(x) = \sum_{i=1}^N f_i(x; \theta_i)$

Stochastic gradients

Access to f provided by *stochastic gradients* of the form

$$G(x; \omega) = \nabla f(x) + Z(x; \omega) \quad (\text{SG})$$

minibatch sampling, ...

Stochastic gradient descent (SGD)

(Robbins & Monro, 1951)

$$x_{t+1} = x_t - \eta_t G(x_t; \omega_t) \quad (\text{SGD})$$

θ_t : i.i.d. sequence of random seeds

η_t : learning rate or step-size (more later)

Background knowledge

What is the long-run behavior of SGD?

Well known

✎ f convex + iterate averaging	$\implies \mathcal{O}(1/\sqrt{T})$ value convergence rate	# near-optimal
✎ f non-convex + random stopping	$\implies \mathcal{O}(1/T^{1/4})$ gradient norm rate	# Ghadimi & Lan (2013)
✎ In both cases:	η_t chosen as a function of T	# ϵ accuracy if T chosen appropriately

Background knowledge

What is the long-run behavior of SGD?

Well known

✎ f convex + iterate averaging	$\implies \mathcal{O}(1/\sqrt{T})$ value convergence rate	# near-optimal
✎ f non-convex + random stopping	$\implies \mathcal{O}(1/T^{1/4})$ gradient norm rate	# Ghadimi & Lan (2013)
✎ In both cases:	η_t chosen as a function of T	# ϵ accuracy if T chosen appropriately
✎ In practice:	Pick some value for η and hope for the best...	# heuristics galore!

Background knowledge

What is the long-run behavior of SGD?

Key challenges

- 👉 *Does SGD converge in the long-run—and in what sense?*
- 👉 *Does SGD avoid non-critical / non-minimizing regions?*
- 👉 *Which points are more likely to be observed in the long run—and by how much?*
- 👉 *How long does it take to reach the vicinity of a global minimum of f ?*

Blanket assumptions

Assumptions for the objective

$f: \mathbb{R}^n \rightarrow \mathbb{R}$

- ▶ **Lipschitz smoothness:** $\|\nabla f(x') - \nabla f(x)\| \leq L\|x' - x\|$ for all $x, x' \in \mathcal{X}$
- ▶ **Coercivity:** $f(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$ # or norm-coercive gradient: $\|\nabla f(x)\| \rightarrow \infty$
- ▶ **Critical set regularity:** $\text{crit}(f)$ is a union of smoothly connected components $\mathcal{K}_i, i = 1, \dots, K$

Blanket assumptions

Assumptions for the objective

$f: \mathbb{R}^n \rightarrow \mathbb{R}$

- ▶ **Lipschitz smoothness:** $\|\nabla f(x') - \nabla f(x)\| \leq L\|x' - x\|$ for all $x, x' \in \mathcal{X}$
- ▶ **Coercivity:** $f(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$ # or norm-coercive gradient: $\|\nabla f(x)\| \rightarrow \infty$
- ▶ **Critical set regularity:** $\text{crit}(f)$ is a union of smoothly connected components $\mathcal{K}_i, i = 1, \dots, K$

Assumptions for the oracle

$G(x; \theta) = \nabla f(x) + Z(x; \theta)$

- ▶ **Properness:** $\mathbb{E}[Z(x; \omega)] = 0$ and $\text{cov}(Z(x; \omega)) > 0$ for all $x \in \mathcal{X}$
- ▶ **Smooth growth:** Z is C^2 -smooth and grows as $\|Z(x; \omega)\| \leq C(1 + \|x\|)$
- ▶ **Sub-Gaussian tails:** $\mathbb{P}(\|Z(x; \omega)\| > z) \leq 2 \exp\left(-\frac{z^2}{2\sigma^2}\right)$

can be relaxed

Setting the step-size

Step-size schedules

How to set the method's step-size?

- ▶ **Vanishing:** $\eta_t \searrow 0$ as $t \rightarrow \infty$ # conservative, more stable
- ▶ **Constant:** $\eta_t \equiv \eta$ for all $t = 1, 2, \dots$ # more aggressive, less stable
- ▶ **Heuristic:** cosine schedules, mix'n'match, adaptive,... # AdaGrad, Adam, ...

Setting the step-size

Step-size schedules

How to set the method's step-size?

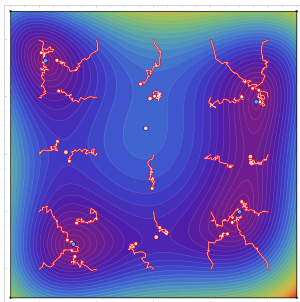
- ▶ **Vanishing:** $\eta_t \searrow 0$ as $t \rightarrow \infty$ # Part I
- ▶ **Constant:** $\eta_t \equiv \eta$ for all $t = 1, 2, \dots$ # Part II
- ▶ **Heuristic:** cosine schedules, mix'n'match, adaptive,... # AdaGrad, Adam, ...

Outline

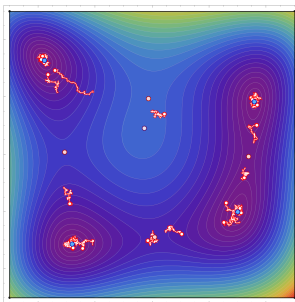
- ① Background & prelims
- ② Part I: SGD with a vanishing step-size
- ③ Part II: SGD with a constant step-size

SGD with a vanishing step-size

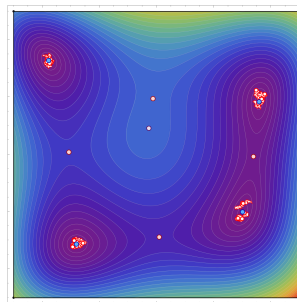
Running (SGD) with $\eta_t \propto 1/\sqrt{t}$



$t = 10$



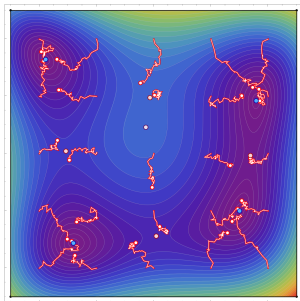
$t = 50$



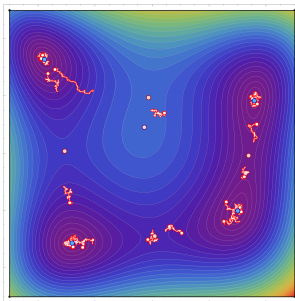
$t = 200$

SGD with a vanishing step-size

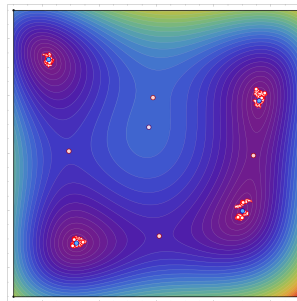
Running (SGD) with $\eta_t \propto 1/\sqrt{t}$



$t = 10$



$t = 50$



$t = 200$

“(SGD) with a vanishing step-size converges to minimizers (and only minimizers)”

From algorithms to flows

Characteristic property of (SGD)

$$\frac{x_{t+1} - x_t}{\eta_t} = -\nabla f(x_t) + Z_t \approx -\nabla f(x_t) \quad \text{“on average”}$$

Gradient flow

$$\dot{x}(t) = -\nabla f(x(t)) \quad (\text{GF})$$

From algorithms to flows

Characteristic property of (SGD)

$$\frac{x_{t+1} - x_t}{\eta_t} = -\nabla f(x_t) + Z_t \approx -\nabla f(x_t) \quad \text{“on average”}$$

Gradient flow

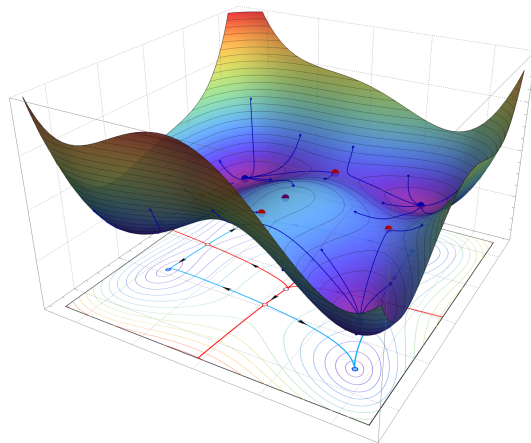
$$\dot{x}(t) = -\nabla f(x(t)) \quad (\text{GF})$$

Basic idea: If η_t is “small”, the errors wash out and “ $\lim_{t \rightarrow \infty} (\text{SGD}) = \lim_{t \rightarrow \infty} (\text{GF})$ ”

Convergence of gradient flows

Lyapunov property of (GF)

$$df/dt = -\|\nabla f(x(t))\|^2 \leq 0 \quad \text{w/ equality iff } \nabla f(x) = 0$$



Convergence of trajectories

Context-specific assumptions:

(A) $\nabla f(x)$ is **subcoercive**:

$$\langle \nabla f(x), x \rangle \geq 0 \quad \text{for sufficiently large } x$$

(B) The parameters of (SGD) satisfy:

▶ $\sum_t \eta_t = \infty$

no stopping

▶ $\eta_t = \mathcal{O}(1/(\log t)^{1+\varepsilon})$ for some $\varepsilon > 0$

more generally: $\sum_t \lambda^{1/\eta_t} < \infty$ for all $\lambda \in (0, 1)$

Theorem (Bertsekas & Tsitsiklis, 2000; M, Hallak, Kavis & Cevher, 2020)

👉 **Assume:** (A) + (B)

👉 **Then:** x_t converges to a connected component of $\mathcal{X}^*$ with probability 1

Are all critical points equally likely?

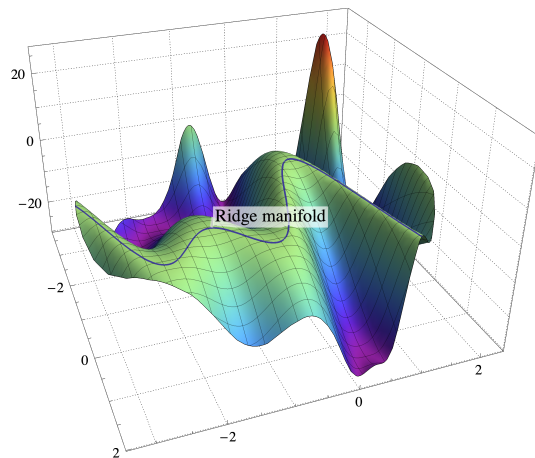


Figure: A hyperbolic ridge manifold, typical of ResNet loss landscapes (Li et al., 2018)

Escaping saddle points

Strict saddles

$$\lambda_{\min}(\nabla^2 f(x^*)) < 0$$

Escaping saddle points

Strict saddles

$$\lambda_{\min}(\nabla^2 f(x^*)) < 0$$

Theorem (Ge et al., 2015)

Given: tolerance level $\zeta > 0$

Assume:

- ▶ f is bounded and Lipschitz smooth
- ▶ $\nabla^2(f(x))$ is Lipschitz continuous
- ▶ One of the following holds for all $x \in \mathcal{X}$:
 - (a) $\|\nabla f(x)\| \geq \varepsilon$
 - (b) $\lambda_{\min}(\nabla^2(f(x))) \leq -\beta$
 - (c) x is δ -close to a local minimum x^* of f around which f is α -strongly convex
- ▶ $Z(x; \omega)$ contains a component uniformly sampled from the unit sphere
- ▶ $\eta_t \equiv \eta$ with $\eta = \mathcal{O}(1/\log(1/\zeta))$

Then: with probability at least $1 - \zeta$, SGD produces after $\mathcal{O}(\eta^{-2} \log(1/(\eta\zeta)))$ iterations a point which is $\mathcal{O}(\sqrt{\eta} \log(1/(\eta\zeta)))$ -close to x^*

Saddle-point avoidance

Theorem (M, Hallak, Kavis & Cevher, 2020)

Assume:

- ▶ The gradient noise is **uniformly exciting**

$$\mathbb{E}[\langle Z(x; \omega), u \rangle^+] \geq c \quad \text{for all unit vectors } u \in \mathbb{S}^{n-1}, x \in \mathcal{X}$$

- ▶ $\eta_t \propto 1/t^p$ for some $p \in (0, 1]$

Then: $\mathbb{P}(x_t \text{ converges to a set of strict saddle points}) = 0$

Saddle-point avoidance

Theorem (M, Hallak, Kavis & Cevher, 2020)

Assume:

- ▶ The gradient noise is **uniformly exciting**

$$\mathbb{E}[\langle Z(x; \omega), u \rangle^+] \geq c \quad \text{for all unit vectors } u \in \mathbb{S}^{n-1}, x \in \mathcal{X}$$

- ▶ $\eta_t \propto 1/t^p$ for some $p \in (0, 1]$

Then: $\mathbb{P}(x_t \text{ converges to a set of strict saddle points}) = 0$

Take-aways:

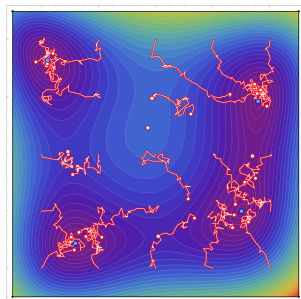
- ▶ **The good:** SGD only converges to local minimizers (or components thereof)
- ▶ **The bad:** only strict saddle points are excluded (degeneracies difficult to deal with)
- ▶ **The ugly:** no known way to predict which minimizers will be selected

Outline

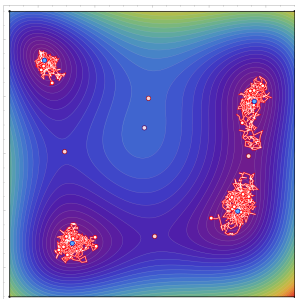
- ① Background & prelims
- ② Part I: SGD with a vanishing step-size
- ③ Part II: SGD with a constant step-size

SGD with a constant step-size

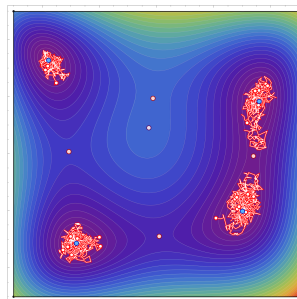
Running (SGD) with $\eta_t \equiv \text{const}$



$t = 10$



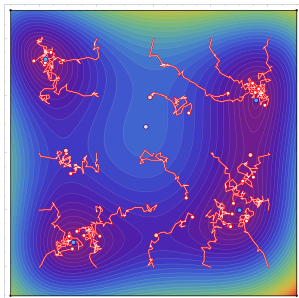
$t = 50$



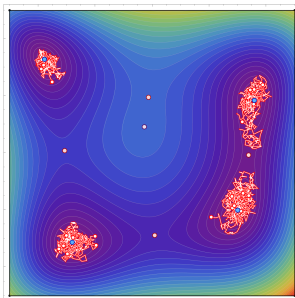
$t = 500$

SGD with a constant step-size

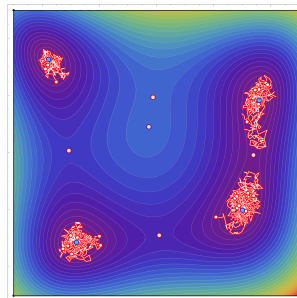
Running (SGD) with $\eta_t \equiv \text{const}$



$t = 10$



$t = 50$



$t = 500$

“(SGD) with a constant step-size ~~converges to~~ concentrates around minimizers (and only minimizers)”

Concentration measures

Occupancy measure

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

average time spent by x_t in $\mathcal{B}$

Concentration measures

Occupancy measure

$$\mu_t(\mathcal{B}) = \mathbb{E} \left[\frac{1}{t} \sum_{\tau=1}^t \mathbb{1}_{\{x_\tau \in \mathcal{B}\}} \right]$$

average time spent by x_t in $\mathcal{B}$

Invariant measure

$$x \sim \mu_\infty \implies x_t \sim \mu_\infty$$

long-run distribution of (SGD)

Concentration measures

Occupancy measure

$$\mu_t(\mathcal{B}) = \mathbb{E} \left[\frac{1}{t} \sum_{\tau=1}^t \mathbb{1}_{\{x_\tau \in \mathcal{B}\}} \right]$$

average time spent by x_t in $\mathcal{B}$

Invariant measure

$$x \sim \mu_\infty \implies x_t \sim \mu_\infty$$

long-run distribution of (SGD)

Goal: Understand the concentration properties of μ_∞

Large deviation principle

Cumulant generating functions

[gradient noise statistics]

$$K_Z(x, p) := \log \mathbb{E} \left[e^{\langle p, Z(x; \omega) \rangle} \right]$$

$$K_G(x, p) := \log \mathbb{E} \left[e^{\langle p, G(x; \omega) \rangle} \right] = \langle \nabla f(x), p \rangle + K_Z(x, p)$$

Large deviation principle

Cumulant generating functions

[gradient noise statistics]

$$K_Z(x, p) := \log \mathbb{E} \left[e^{\langle p, Z(x; \omega) \rangle} \right]$$

$$K_G(x, p) := \log \mathbb{E} \left[e^{\langle p, G(x; \omega) \rangle} \right] = \langle \nabla f(x), p \rangle + K_Z(x, p)$$

Lagrangians

[distance to $-\nabla f(x)$]

$$\mathcal{L}_Z(x, v) := K_Z^*(x, -v)$$

$$\mathcal{L}_G(x, v) := K_G^*(x, -v) = \mathcal{L}_Z(x, v + \nabla f(x))$$

NB: “ $*$ ” $\rightsquigarrow$ convex conjugation relative to p

Large deviation principle

Cumulant generating functions

[gradient noise statistics]

$$K_Z(x, p) := \log \mathbb{E} \left[e^{\langle p, Z(x; \omega) \rangle} \right]$$

$$K_G(x, p) := \log \mathbb{E} \left[e^{\langle p, G(x; \omega) \rangle} \right] = \langle \nabla f(x), p \rangle + K_Z(x, p)$$

Lagrangians

[distance to $-\nabla f(x)$]

$$\mathcal{L}_Z(x, v) := K_Z^*(x, -v)$$

$$\mathcal{L}_G(x, v) := K_G^*(x, -v) = \mathcal{L}_Z(x, v + \nabla f(x))$$

NB: “ $*$ ” $\rightsquigarrow$ convex conjugation relative to p

Large deviation principle

[Cramér's theorem]

$$\mathbb{P} \left(\frac{1}{t} \sum_{\tau=1}^t G(x; \omega_\tau) \in \mathcal{B} \right) \sim \exp \left(-t \inf_{v \in \mathcal{B}} \mathcal{L}_G(x, v) \right)$$

Freidlin-Wentzell theory

Seek: Large deviation principle for $x_t = \eta \sum_{\tau=1}^{t-1} G(\mathbf{x}_\tau; \omega_\tau)$ instead of $\sum_{\tau=1}^{t-1} G(\mathbf{x}; \omega_\tau)$

Freidlin-Wentzell theory

Seek: Large deviation principle for $x_t = \eta \sum_{\tau=1}^{t-1} G(\mathbf{x}_\tau; \omega_\tau)$ instead of $\sum_{\tau=1}^{t-1} G(\mathbf{x}; \omega_\tau)$

Action functional

$$\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}_G(\gamma(t), \dot{\gamma}(t)) dt$$

Basic element of Freidlin-Wentzell theory (Freidlin & Wentzell, 1998)

Freidlin-Wentzell theory

Seek: Large deviation principle for $x_t = \eta \sum_{\tau=1}^{t-1} G(\mathbf{x}_\tau; \omega_\tau)$ instead of $\sum_{\tau=1}^{t-1} G(\mathbf{x}; \omega_\tau)$

Action functional

$$\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}_G(\gamma(t), \dot{\gamma}(t)) dt$$

Basic element of Freidlin-Wentzell theory (Freidlin & Wentzell, 1998)

Proposition (Azizian, Iutzeler, Malick & M, 2024; informal)

“Trajectories of (SGD) tend to concentrate near action-minimizing curves”

Discrete-time version of Freidlin-Wentzell LDP

Transition costs

How probable is a transition from one critical region to another?

Transition costs

How probable is a transition from one critical region to another?

Transition costs

(Kifer, 1990)

- ▶ **Quasi-potential:** $B(x, x') := \inf\{\mathcal{S}_T[\gamma] : \gamma(0) = x, \gamma(T) = x', T \in \mathbb{N}\}$ # least action path joining $x \rightsquigarrow x'$
- ▶ **Transition cost:** $B(\mathcal{K}, \mathcal{K}') = \inf\{B(x, x') : x \in \mathcal{K}, x' \in \mathcal{K}'\}$ # least action joining $\mathcal{K} \rightsquigarrow \mathcal{K}'$
- ▶ **Cost matrix:** $B_{ij} = B(\mathcal{K}_i, \mathcal{K}_j)$ # transition cost from $\mathcal{K}_i$ to $\mathcal{K}_j$

Transition costs

How probable is a transition from one critical region to another?

Transition costs

(Kifer, 1990)

- ▶ **Quasi-potential:** $B(x, x') := \inf\{\mathcal{S}_T[\gamma] : \gamma(0) = x, \gamma(T) = x', T \in \mathbb{N}\}$ # least action path joining $x \rightsquigarrow x'$
- ▶ **Transition cost:** $B(\mathcal{K}, \mathcal{K}') = \inf\{B(x, x') : x \in \mathcal{K}, x' \in \mathcal{K}'\}$ # least action joining $\mathcal{K} \rightsquigarrow \mathcal{K}'$
- ▶ **Cost matrix:** $B_{ij} = B(\mathcal{K}_i, \mathcal{K}_j)$ # transition cost from $\mathcal{K}_i$ to $\mathcal{K}_j$

Transition graph of SGD

Complete weighted directed graph $\mathcal{G} \equiv \mathcal{G}(\mathcal{V}, \mathcal{E})$ with primitives:

- ▶ **Vertices:** $\mathcal{V} = \{1, \dots, K\}$ # critical components
- ▶ **Edges:** $\mathcal{E} = \{(i, j) : i, j = 1, \dots, K\}$ # transitions between components
- ▶ **Weights:** $(i, j) \rightsquigarrow B_{ij}$ # transition costs

The transition graph

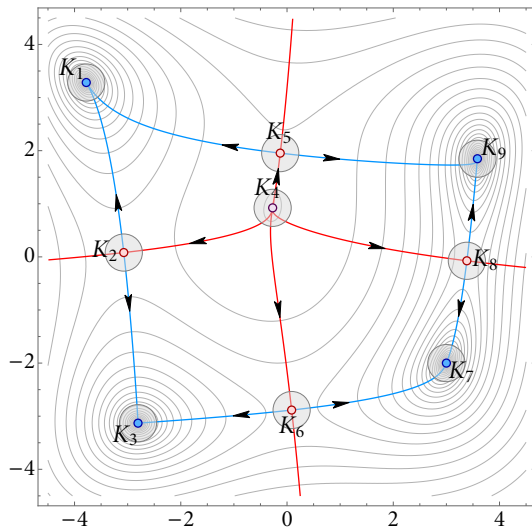


Figure: Transition graph of (SGD) for the Himmelblau function

Energy, Gibbs, and the long-run distribution of SGD

The transition graph defines the *energy* of each component:

$$\mathcal{E}_i = \min_{T_i \in \mathcal{T}_i} \sum_{j, k \in T_i} B_{jk}$$

minimization over spanning trees $T_i \in \mathcal{T}_i$ pointing to i

Energy, Gibbs, and the long-run distribution of SGD

The transition graph defines the **energy** of each component:

$$\mathcal{E}_i = \min_{T_i \in \mathcal{T}_i} \sum_{j, k \in T_i} B_{jk}$$

minimization over spanning trees $T_i \in \mathcal{T}_i$ pointing to i

Theorem (Azizian, Iutzeler, Malick & M, 2024; informal)

1. The critical region of f is visited exponentially more often than any non-critical region
2. Critical components are visited according to the **Boltzmann-Gibbs** distribution with temperature η
3. **Minimizing components** are visited exponentially more often than **non-minimizing** ones
4. The **ground state** of the problem is visited exponentially more often than any other component

Energy, Gibbs, and the long-run distribution of SGD

The transition graph defines the **energy** of each component:

$$\mathcal{E}_i = \min_{T_i \in \mathcal{T}_i} \sum_{j, k \in T_i} B_{jk}$$

minimization over spanning trees $T_i \in \mathcal{T}_i$ pointing to i

Theorem (Azizian, Iutzeler, Malick & M, 2024)

Fix a tolerance $\varepsilon > 0$ and let $\mathcal{U}_i \equiv \mathcal{U}_i(\delta)$ be a δ -neighborhood of $\mathcal{K}_i$. Then, for all sufficiently small $\delta, \eta > 0$, we have:

1. If $\mathcal{U} = \bigcup_{i=1}^K \mathcal{U}_i$ is a δ -neighborhood of $\text{crit}(f)$, then $\mu_\infty(\mathcal{U}) \geq 1 - e^{-c/\eta}$ for some $c > 0$
2. For all $i = 1, \dots, K$

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

3. If $\mathcal{K}_i$ is **non-minimizing**, there exists a **minimizing** component $\mathcal{K}_j$ such that $c_{ij} := \mathcal{E}_i - \mathcal{E}_j > 0$, and hence

$$\frac{\mu_\infty(\mathcal{U}_i)}{\mu_\infty(\mathcal{U}_j)} \leq \exp\left(-\frac{c_{ij} + \mathcal{O}(\varepsilon)}{\eta}\right)$$

4. If $\mathcal{U}_0$ is a δ -neighborhood of the **ground state** $\mathcal{K}_0 = \arg \min_i \mathcal{K}_i$, then $\mu_\infty(\mathcal{U}_0) \geq 1 - e^{-c_0/\eta}$ for some $c_0 > 0$

Energy landscape

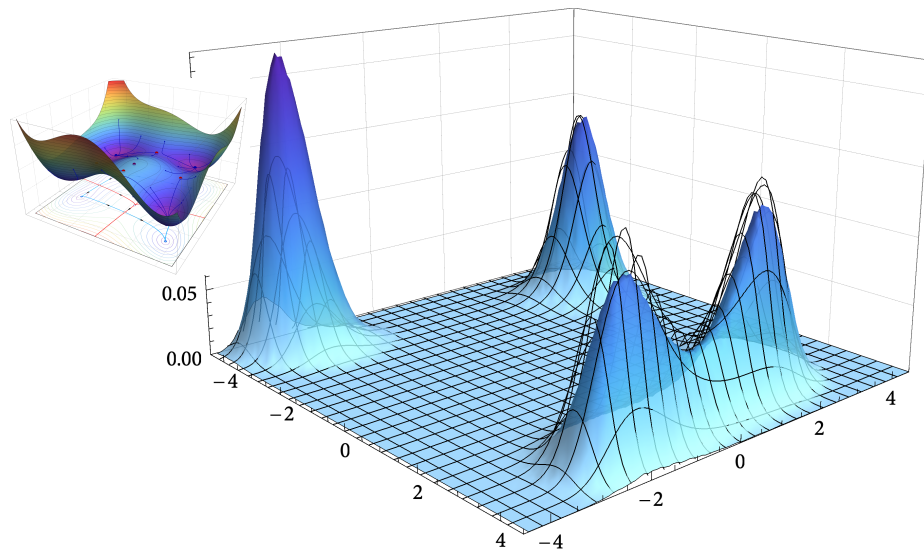
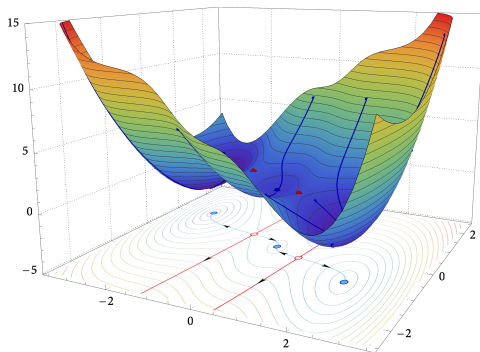


Figure: Energy landscape of the Himmelblau function (wireframe) and empirical distribution of SGD (spacefill)

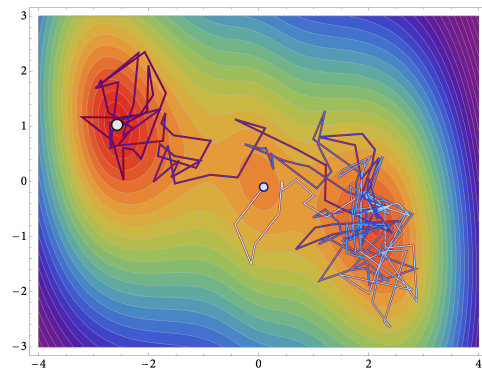
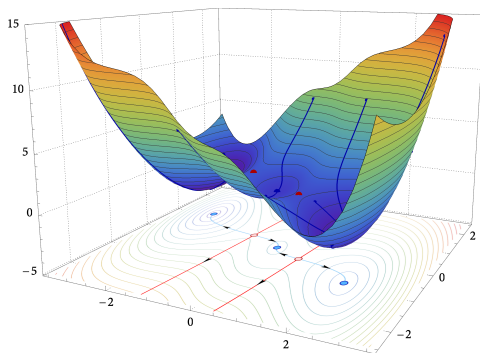
Reaching the global minimum

What is the average time required to reach a neighborhood of $\arg \min f$?



Reaching the global minimum

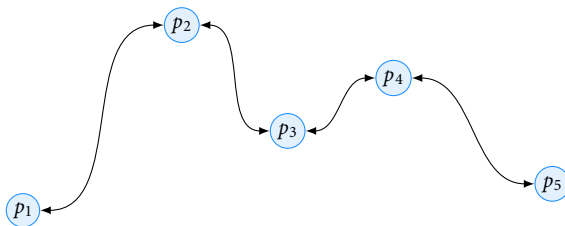
What is the average time required to reach a neighborhood of $\arg \min f$?



The global convergence time of SGD

Global convergence time

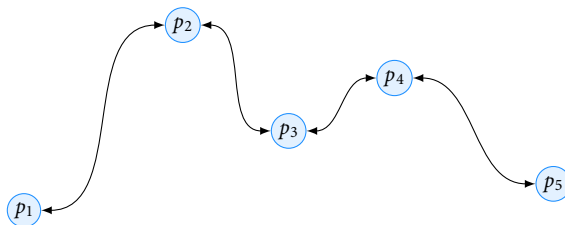
$$\tau = \inf \{t \geq 0 : \text{dist}(\arg \min f, x_t) \leq \delta\}$$



The global convergence time of SGD

Global convergence time

$$\tau = \inf \{ t \geq 0 : \text{dist}(\arg \min f, x_t) \leq \delta \}$$



Theorem (Azizian et al, 2025; applied to the above example)

Suppose (SGD) is initialized near p_3 . Then:

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

Relative energy and the global convergence time of SGD

Energy of $\mathcal{Q} = \arg \min f$ relative to a starting node p :

$$\mathcal{E}[\mathcal{Q} | p] = \mathcal{E}[\mathcal{Q}] - \min_{S \in \mathcal{G}(p \not\rightarrow \mathcal{Q})} \sum_{(i,j) \in S} B_{ij}.$$

Theorem (Azizian et al, 2025)

Given a tolerance level $\varepsilon > 0$, an initialization $x \in \mathbb{R}^n$ of (SGD), and small enough $\delta, \eta > 0$, we have

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

Relative energy and the global convergence time of SGD

Energy of $\mathcal{Q} = \arg \min f$ relative to a starting node p :

$$\mathcal{E}[\mathcal{Q} | p] = \mathcal{E}[\mathcal{Q}] - \min_{\mathcal{S} \in \mathcal{G}(p \not\rightsquigarrow \mathcal{Q})} \sum_{(i,j) \in \mathcal{S}} B_{ij}.$$

Theorem (Azizian et al, 2025)

Given a tolerance level $\varepsilon > 0$, an initialization $x \in \mathbb{R}^n$ of (SGD), and small enough $\delta, \eta > 0$, we have

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

Core Take-Away

$$\mathcal{E}[\mathcal{Q} | x] \leq 2 \left[\frac{\# \{\text{spurious minimizers}\} \times \# \{\text{max. saddle - min. spurious min.}\}}{\sigma_+^2} - \frac{\text{depth of global minimum}}{\sigma_0^2} \right]$$

where

- ▶ σ_0^2 = variance of G near $\arg \min f$
- ▶ σ_+^2 = infimum of the variance of G near spurious minimizers

Take-aways and conclusions

I. SGD with vanishing step-size:

- ▶ With probability 1 every trajectory converges to a component of critical points of f
- ▶ Components of strict saddle-points are avoided with probability 1

II. SGD with a constant step-size:

- ▶ Critical regions are visited exponentially more often than non-critical regions
- ▶ **Minimizing components** are visited exponentially more often than **non-minimizing** ones
- ▶ Critical components are visited with probability exponentially proportional to their **energy**
! not necessarily the value of f !
- ▶ The **ground state** of the problem is visited exponentially more often than any other component
! not necessarily $\arg \min f$!

Many implications but also many open questions:

- ▶ Other step-size policies? # AdaGrad-like, periodic...
- ▶ Other gradient-based algorithms? # Adam?
- ▶ Beyond minimization?

References

- [1] Azizian, W., Iutzeler, F., Malick, J., and Mertikopoulos, P. What is the long-run distribution of stochastic gradient descent? A large deviations analysis. In *ICML '24: Proceedings of the 41st International Conference on Machine Learning*, 2024.
- [2] Azizian, W., Iutzeler, F., Malick, J., and Mertikopoulos, P. The global convergence time of stochastic gradient descent in non-convex landscapes: Sharp bounds via large deviations. In *ICML '25: Proceedings of the 42nd International Conference on Machine Learning*, 2025.
- [3] Bertsekas, D. P. and Tsitsiklis, J. N. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3): 627–642, 2000.
- [4] Freidlin, M. I. and Wentzell, A. D. *Random Perturbations of Dynamical Systems*. Springer, 1 edition, 1998.
- [5] Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points - Online stochastic gradient for tensor decomposition. In *COLT '15: Proceedings of the 28th Annual Conference on Learning Theory*, 2015.
- [6] Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- [7] Kifer, Y. A discrete-time version of the Wentzell-Freidlin theory. *Annals of Probability*, 18(4):1676–1692, October 1990.
- [8] Li, H., Xu, Z., Taylor, G., Suder, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.
- [9] Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. On the almost sure convergence of stochastic gradient descent in non-convex problems. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.
- [10] Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.