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

## A Large Deviation Analysis

**Franck IUTZELER**
Université de Toulouse

JS2O – Perpignan – April 3rd, 2025

# Motivation • Neural networks' complicated landscape

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

Image credit: losslandscape.com

- Objective function $f : \mathbb{R}^d \to \mathbb{R}$ smooth nonconvex
- Gradient Descent

$$x_{n+1} = x_n - \underbrace{\eta}_{\text{stepsize}} \nabla f(x_n)$$

  - Converges to some critical point depending solely on initialization
  - If stepsize $\eta$ is small: close to gradient flow $\dot{X}_t = -\nabla f(X_t)$
  - Needs a full gradient evaluation $\rightsquigarrow$ out-of-reach in large-scale learning

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

Stochastic first-order oracle = what is computed

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

stepsize | zero-mean noise

  - Converges to some local minimum depending on initialization and noise
  - Asymptotically close to gradient flow $\dot{X}_t = -\nabla f(X_t)$ with probability one
  - Can get trapped in local minima

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

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

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

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

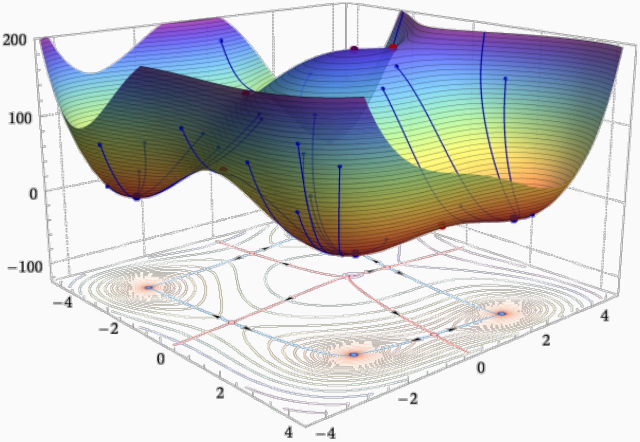
$$x_{n+1} = x_n - \eta \left[ \nabla f(x_n) + Z(x_n; \omega_{n+1}) \right]$$

stepsize ($\eta$) — zero-mean noise ($Z(x_n; \omega_{n+1})$)

  - No pointwise convergence, the exploration due to the noise does not vanish
  - Very efficient in practice for machine learning problems

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

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



- Lines of work that **do not characterize the asymptotic behavior**
  - Sampling (MCMC, Langevin) scaling of the noise differs from SGD

    $$x_{n+1} = x_n - \eta \nabla f(x_n) + \sqrt{2\eta}\, \xi_n \quad \text{with } \xi_n \sim \mathcal{N}(0, \sigma^2)$$

  - Continuous-time limit (SDE) only valid on finite time horizons [Li et al., 2017]

    $$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2\eta\, \mathrm{cov}(Z(X_t;\cdot))}\,\mathrm{d}W_t$$

- **Classical results** in optimization
  - 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]
  - 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
  - arXiv 2503.16398 Fresh out!

# Setup & Assumptions

- Objective function $f$
  - smooth — $C^2$ and $\nabla f$ is $\beta$-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 — $Z$ is $C^2$ and $Z(x;\omega) = \mathcal{O}(\|x\|)$ almost surely
  - sub-Gaussian — $\log \mathbb{E}[\exp(\langle p, Z(x;\omega)\rangle)] \le \frac{\sigma^2\|p\|^2}{2}$

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 $\nabla f$ is $\beta$-Lipschitz continuous
  - coercive — $\lim_{\|x\|\to\infty} f(x) = +\infty$
  - gradient coercive — $\lim_{\|x\|\to\infty} \|\nabla f(x)\| = +\infty$

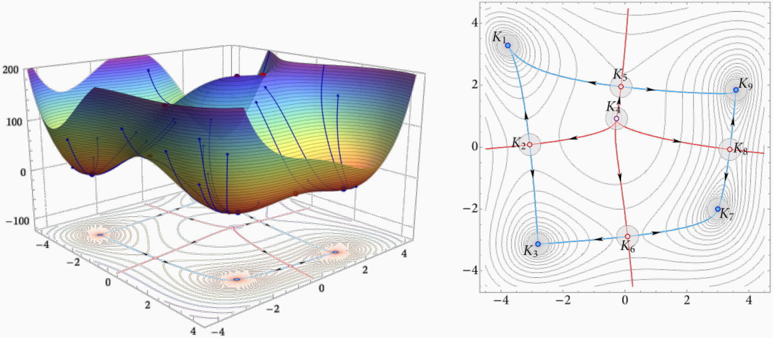
- Noise term $Z$
  - proper — $\mathbb{E}[Z(x;\omega)] = 0$ and $\mathrm{cov}(Z(x;\omega)) \succ 0$ for all $x \in \mathbb{R}^d$
  - limited growth — $Z$ is $C^2$ and $Z(x;\omega) = \mathcal{O}(\|x\|)$ almost surely
  - sub-Gaussian — $\log \mathbb{E}[\exp(\langle p, Z(x;\omega)\rangle)] \leq \frac{\sigma^2\|p\|^2}{2}$

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

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

where $\omega$ 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** $\mu_\infty^\eta$ 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 $\mu_\infty^\eta(\mathcal{K}_i)/\mu_\infty^\eta(\mathcal{K}_j)$

# 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 + Lyapunov condition $\Rightarrow \exists$ 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 (typically $\mathcal{L}(x,v) \geq 0$ and $\mathcal{L}(x,v) = 0 \iff v = -\nabla f(x)$)

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

- Markov chain
  - weak Feller + Lyapunov condition $\Rightarrow \exists$ 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 (typically $\mathcal{L}(x,v) \geq 0$ and $\mathcal{L}(x,v) = 0 \iff v = -\nabla f(x)$)

# Result 1 • a LDP for SGD

- **Ingredients** for comparing SGD w/ a smooth curve $\gamma : [0,T] \to \mathbb{R}^d$
  - **Cumulant Generating Function** $K(x,p) \coloneqq \log \mathbb{E}[\exp(\langle p, Z(x;\omega)\rangle)] + \langle \nabla f(x), p\rangle$
  - **Lagrangian** $\mathcal{L}(x,v) \coloneqq K^*(x,-v)$ is its convex conjugate (in $v$)
  - **Action functional** $\mathcal{S}_T[\gamma] = \int_0^T \mathcal{L}(\gamma(t),\dot\gamma(t))\,dt$

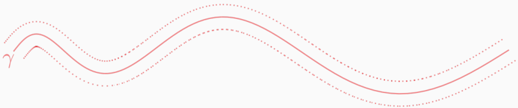
**Result 1** As $\eta \to 0$

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

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

**Gaussian case** $\mathcal{L}(x,v) = \frac{\|v+\nabla f(x)\|^2}{2\sigma^2}$

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



- Discrete time

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

- 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 $Z_0 = x_0$

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

$\mathrm{dist}_{[0,T]}(X, Z) \leq c\eta$ for some $c$

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 $\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2\eta\,\mathrm{cov}(\mathsf{Z}(X_t; \cdot))}\,\mathrm{d}W_t$ is different, has the wrong scale for the noise, and the discretization or the convergence is exponentially bad in $\eta$ [ Raginsky et al., 2017 ; Li et al., 2019]

**Proposition** As $\eta \to 0$,

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

- **Idea** inspired from [Freidlin and Wentzell, 1998]
  - $\{0, 1/\eta, .., T/\eta\}$ iterates of SGD $\approx [0,T]$ trajectory of $\dot{Z}_t = -\nabla f(Z_t) + \mathrm{Z}(Z_t, \omega_{\lfloor t/\eta \rfloor})$
  - **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 $\gamma \in C_T$

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

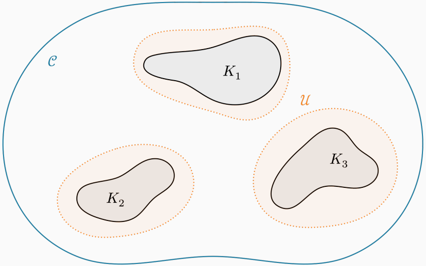
**Result 1** As $\eta \to 0$

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

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

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

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



# Transitions between critical components

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

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

  - **fixes** some transition **time** $T$
  - if there is a **gradient flow** going from $x$ to $x'$, then $B(x, x') = 0$

- Potentials for transitioning **between critical components**

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

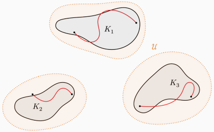
  - From **Result 1**, we have for $\eta > 0$ small enough

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

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

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

  - Transitions probabilities are given by the $B_{ij}$




**Result 2** The invariant distribution $\pi$ 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 $i$

- From Result 1, critical neighborhoods are exponentially more visited so the invariant distribution of $z_n$ captures the long-run behavior of SGD

# 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.

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

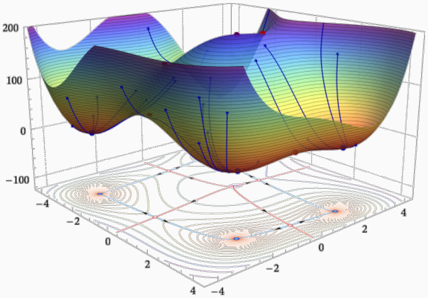
- **Boltzmann-Gibbs distribution** for all $i$

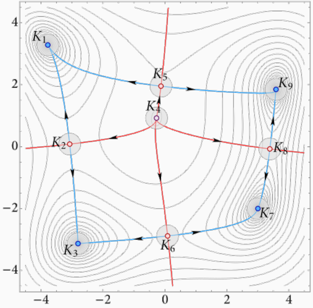
$$\mu_\infty^\eta(\mathcal{U}_i) \propto \exp\left(-\frac{E_i + \mathcal{O}(\varepsilon)}{\eta}\right)$$

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

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

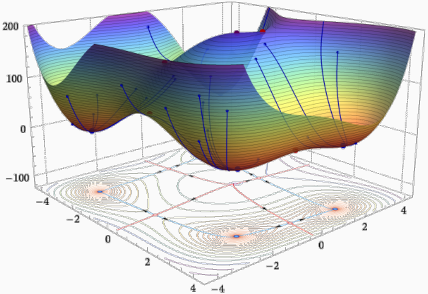
- Assume that $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I)$
  - $B_{51} = 0 \quad B_{15} = 2(f(x_5) - f(x_1))/\sigma^2$ for $(x_1, x_5) \in \mathcal{K}_1 \times \mathcal{K}_5$
  - $E_i = 2f(x_i)/\sigma^2$ for any $x_i \in \mathcal{K}_i$

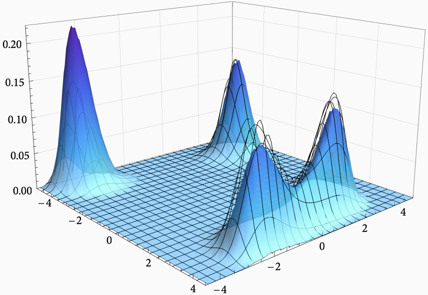


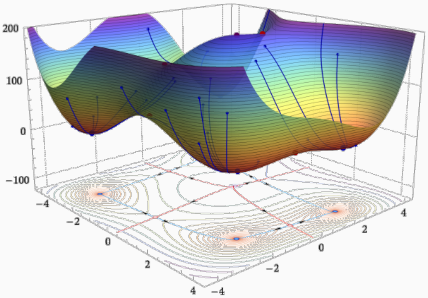
# Example • Himmelblau with Gaussian noise

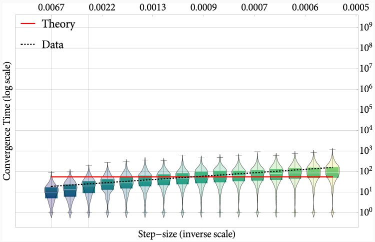
- Assume that $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I)$
  - $B_{51} = 0 \quad B_{15} = 2(f(x_5) - f(x_1))/\sigma^2$ for $(x_1, x_5) \in \mathcal{K}_1 \times \mathcal{K}_5$
  - $E_i = 2f(x_i)/\sigma^2$ for any $x_i \in \mathcal{K}_i$



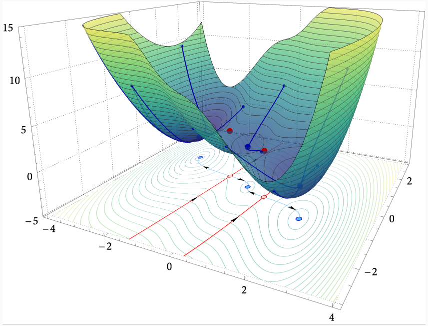



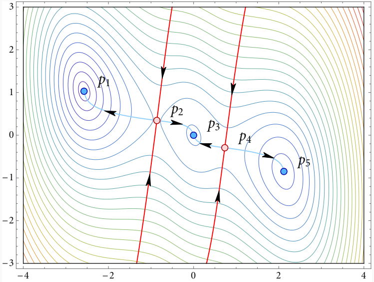
- Building on the **transition probabilities** before, we can characterize the **average time to reach the global minimum**
- **Himmelblau** function: 9 critical points, 4 global min
- Assume as before that $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I)$
  - We can show that $\mathbb{E}[$**time to reach a global min**$]$ **does not depend on** $\eta$



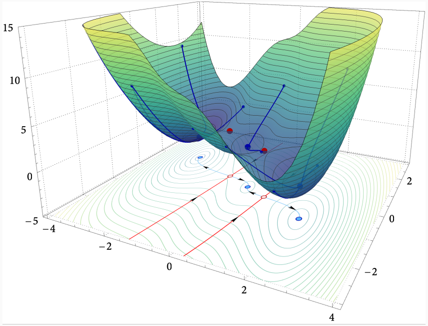



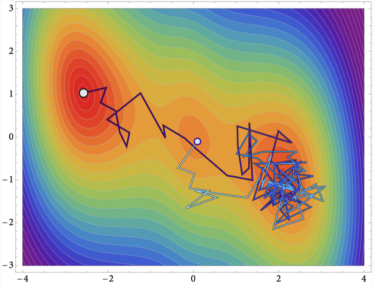
- Building on the **transition probabilities** before, we can characterize the **average time to reach the global minimum**
- **Three-humps camel** function: 5 critical points ($p_i$), $p_1$ is the global min
- Assume as before that $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I)$
  - We can show that $\mathbb{E}[\text{time to reach } p_1] \approx \exp(\frac{2(f(p_2)-f(p_5))}{\eta\sigma^2})$





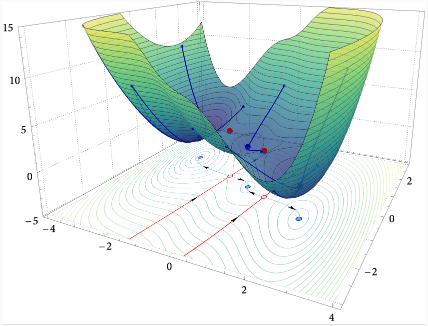
- Building on the **transition probabilities** before, we can characterize the **average time to reach the global minimum**
- **Three-humps camel** function: 5 critical points $(p_i)$, $p_1$ is the global min
- Assume as before that $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I)$
  - We can show that $\mathbb{E}[\text{time to reach } p_1] \approx \exp(\frac{2(f(p_2)-f(p_5))}{\eta\sigma^2})$
  - Here we start near $p_3$ (we can easily go to $p_5$ !)

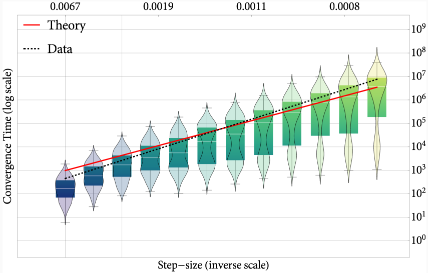


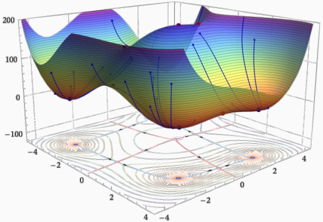
# Going Further • Reaching the global minimum

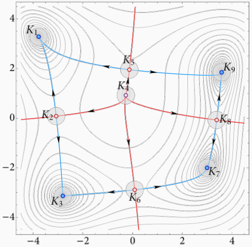
- Building on the **transition probabilities** before, we can characterize the **average time to reach the global minimum**
- **Three-humps camel** function: 5 critical points $(p_i)$, $p_1$ is the global min
- Assume as before that $Z(x;\omega) \sim \mathcal{N}(0, \sigma^2 I)$
  - We can show that $\mathbb{E}[\text{time to reach } p_1] \approx \exp(\frac{2(f(p_2)-f(p_5))}{\eta\sigma^2})$ $\rightsquigarrow$ **good match!**
  - Here we start near $p_3$ (we can easily go to $p_5$ !)

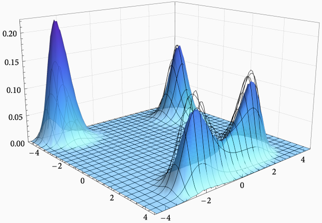


- We introduce a theory of **large deviations** for **SGD in nonconvex problems**
  - Sound approach for the **long-run of SGD**
  - **Precise adaptation** of random perturbations of dynamical systems' theory
- We characterize the **asymptotic distribution of SGD**
  - **Critical regions** are visited **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***

***www.iutzeler.org***