MATHEMATICS OF DEEP LEARNING

JOAN BRUNA , CIMS + CDS, NYU, SPRING'18

Lecture 9: Continuous time in Optimization: SDEs and Landscapes

OBJECTIVES LECTURE 9

- Stochastic Modified Equations

- Stochastic Gradient Langevin Dynamics

- Energy Landscapes and gradient descent: Gradient Descent (stochastic and deterministic) Converges to local Minimisers.

CONTINUOUS-TIME INTERPRETATION

- The error expansion in terms of the step-size can be seen as a discretization error of an underlying continuous (and stochastic) dynamics, given by the stochastic gradient flow

$$\dot{\theta}_t = -\nabla g(\theta_t) + \text{ noise } .$$

- How to deal with that kind of stochastic equation?
- Handle non-convex objectives as well?
- Relationship with other diffusion schemes?

STOCHASTIC DIFFERENTIAL EQUATIONS

- Let $T > 0$. An Ito stochastic differential equation on the interval [0,T] is of the form

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t \ , \ X_0 = x_0 \, , \text{with}$$

$$X_t \in \mathbb{R}^d \, , b : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d \, , \sigma : \mathbb{R}^d \times [0, T] \to \mathbb{R}^{d \times l}$$

$$W_t \text{: } l\text{-dimensional Brownian Motion.}$$

- A Brownian Motion (or Wiener process) W_t is such that:
  - $W_0 = 0$ a.s.
  - W_t has independent and Gaussian increments:
    $W_{t+u} - W_t \sim \mathcal{N}(0, u\mathbf{1})$,
    $W_{t+u} - W_t$ independent of $W_s, s < t$.
  - Sample paths are continuous with probability 1.

ITO CALCULUS

➤ This formalizes the intuition that SDEs are "noisy" ODEs:

$$\dot{\theta}_t = -\nabla g(\theta_t) + \text{ noise } .$$

$$\frac{dX_t}{dt} = b(X_t, t) + \sigma(X_t, t)\frac{dW_t}{dt}$$

➤ The SDE is associated with the stochastic integral equation

$$X_t - x_0 = \int_0^t b(X_s, s)ds + \underbrace{\int_0^t \sigma(X_s, s)dW_s}_{\text{Ito integral}}$$

$$\int_0^t F_s dW_s = \lim_{m\to\infty} \sum_{0=s_0<\cdots<s_m=t} F_{s_{i-1}}(W_{s_i} - W_{s_{i-1}}) .$$

DISCRETIZATION OF SDES

- Consider the SDE

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t \ , \ X_0 = x_0 \, , \text{with}$$

- A stochastic discrete-time scheme $(x_k)_k$ approximates this SDE *in the weak sense* at order α if for every polynomial-growth function g $s.t.$ $|g(x)| \leq K(1 + |x|^\kappa)$ exists C>0 independent of δ such that for all $k = 0, 1, \ldots T/\delta$

$$|\mathbb{E}g(X_{k\delta}) - \mathbb{E}g(x_k)| \leq C\delta^\alpha \, .$$

- Weak approximations do not necessarily have similar sample paths, but rather their distribution.

DISCRETIZATIONS OF SDES: EULER MARUYAMA

- The Euler Maruyama method extends the Euler discretization of ODEs to SDEs.
- Consider as before the SDE

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t \ , \ X_0 = x_0 \, , \text{with}$$

- Fix a time sampling interval $\delta > 0$ and define $x_k := X_{k\delta}$
- Consider the finite difference equation

$$x_{k+1} = x_k + \delta b(x_k, k\delta) + \sigma(x_k, k\delta)(W_{(k+1)\delta} - W_{k\delta}) \, .$$

- Since $W_{(k+1)\delta} - W_{k\delta} \sim \mathcal{N}(0, \delta \mathbf{1})$, this is equivalent to

$$x_{k+1} = x_k + \delta b(x_k, k\delta) + \sqrt{\delta}\sigma(x_k, k\delta)Z_k \ , \ Z_k \sim \mathcal{N}(0, \mathbf{1}) \, .$$

- This scheme is a first order weak approximation to the SDE.

STOCHASTIC MODIFIED EQUATIONS [LI ET AL.'17]

- Consider the Empirical Risk Minimization setup (no reg):

$$\min_{\theta \in \mathbb{R}^d} g(\theta) = \frac{1}{n} \sum_{i \leq n} g_i(\theta)$$

- Stochastic Gradient Descent with fixed step-size:

$$\theta_{k+1} = \theta_k - \gamma \nabla g_{m_k}(\theta_k)\,,$$
$$\{m_k\}\text{: iid uniform variables in } \{1, \dots, n\}.$$

- Goal: derive SDEs that can be seen as weak limits of stochastic gradient descent.
  - In deterministic systems, these are called *modified equations*.

STOCHASTIC MODIFIED EQUATIONS

- We decompose the stochastic gradient in terms of its mean:

$$V_k := \sqrt{\gamma}(\nabla g(\theta_k) - \nabla g_{m_k}(\theta_k))$$

- We have that

$$\theta_{k+1} - \theta_k = -\gamma \nabla g(\theta_k) + \sqrt{\gamma} V_k$$

- Conditioned on θ_k, V_k has mean 0 and covariance $\gamma\Sigma(\theta_k)$:

$$\Sigma(\theta) = \frac{1}{n} \sum_{i \leq n} (\nabla g(\theta) - \nabla g_i(\theta))(\nabla g(\theta) - \nabla g_i(\theta))^\top .$$

- Does this look similar to the Euler-Maruyama scheme for an appropriately chosen SDE?

STOCHASTIC MODIFIED EQUATIONS

- **Theorem [Li, Tai, E, '17]:** Let $T > 0$ and $\Sigma(\theta) \in \mathbb{R}^{d\times d}$ as before. Assuming g, g_k are sufficiently smooth, then

$$dX_t = -\nabla g(X_t)dt + (\gamma\Sigma(X_t))^{1/2}dW_t$$

is an order 1 weak approximation of the SGD.

$$dX_t = -\nabla(g(X_t) + \frac{\gamma}{4}\|\nabla g(X_t)\|^2)dt + (\gamma\Sigma(X_t))^{1/2}dW_t$$

is an order 2 weak approximation of the SGD.

- No convexity assumptions on g.
- How easy is to solve SDEs like above?

SGD DYNAMICS

- We obtained an SDE of the form

$$dX_t^{(\epsilon)} = b(X_t^{(\epsilon)})dt + \epsilon\sigma(X_t^{(\epsilon)})dW_t \ \text{, with } \epsilon \ll 1\,.$$

- Stochastic Asymptotic Expansions [Friedlin et al'12] consider asymptotic expansions wrt ϵ, assuming b and σ are smooth:

$$X_t^{(\epsilon)} = X_{0,t} + \epsilon X_{1,t} + \epsilon^2 X_{2,t} + \dots$$
$$b(X_t^{(\epsilon)}) = b(X_{0,t}) + \epsilon\nabla b(X_{0,t})X_{1,t} + O(\epsilon^2)$$
$$\sigma(X_t^{(\epsilon)}) = \sigma(X_{0,t}) + \epsilon\nabla\sigma(X_{0,t})X_{1,t} + O(\epsilon^2)$$

- Identifying terms with same order we obtain

$$dX_{0,t} = b(X_{0,t})dt$$
$$dX_{1,t} = \nabla b(X_{0,t})X_{1,t}dt + \sigma(X_{0,t})dW_t$$
$$\dots$$

SGD DYNAMICS

- Applying the Stochastic Asymptotic Expansion in the SGD case we obtain

$$X_t \sim \mathcal{N}(X_{0,t}, \gamma S_t) \ ,$$

$$\dot{X}_{0,t} = -\nabla g(X_{0,t}) \quad \dot{S}_t = -S_t H_t - H_t S_t + \Sigma_t$$

$$\text{with } H_t = \nabla^2 g(X_{0,t}), \Sigma_t = \Sigma(X_{0,t}) \ .$$

- This implies that the steady state of S_t satisfies $|S_\infty| \sim \dfrac{|\Sigma_\infty|}{|H_\infty|}$
- Two phases in SGD:
  - Descent regime dominated by gradient flow.
  - Fluctuating regime dominated by S_t.

FROM SDES TO PDES

- Consider an SDE of the form *drift* + *diffusion*:

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t$$

- Denote by $p(x,t),\ x \in \mathbb{R}^d\ , t \in \mathbb{R}_+$ the probability density of X_t. How does $p(x,t)$ evolve with time?

- The *Fokker-Plank equation* is given by

$$\partial_t p(x,t) = -\langle \nabla_x, p(x,t) \cdot b(x,t) \rangle + \frac{1}{2} \langle \Delta_x, p(x,t) \cdot (\sigma\sigma^\top)(x,t) \rangle \ .$$

  - The evolution of the density is deterministic. If SDE models a discrete-time scheme, its associated Fokker-Plank equation models the limit distribution of associated solutions.

EXAMPLE: BROWNIAN MOTION

- Set drift term $b \equiv 0$ and isotropic diffusion $\sigma(x,t) \equiv 2D > 0$.

$$dX_t = 2D dW_t \ : \ \text{Brownian Motion.}$$

- The associated Fokker-Plank equation becomes

$$\partial_t p = D\Delta p = D \sum_{i=1}^{d} \frac{\partial^2 p}{\partial x_i^2}$$

$$\text{Initial condition: } p(x,0) = \delta(x-y)$$

- This is the heat equation, whose solution converges to a Gaussian distribution $\mathcal{N}(y, D\sqrt{t}\mathbf{1})$.

GRADIENT FLOWS

- Consider the SDE

$$dX_t = -\nabla g(X_t)dt + \sqrt{2D}dW_t$$

- and its associated Fokker-Plank equation:

$$\partial_t p = \nabla \cdot (\nabla g p) + D\Delta p$$

- This is a diffusion process with potential function g.
- In general, we cannot explicitly find the time-dependent solution to the FP equation.
- But we can always compute its *stationary* solution.

GRADIENT FLOWS AND FOKKER-PLANK

- **Proposition:** If $g(x)$ is smooth and $e^{-D^{-1}g(x)} \in L^1(\mathbb{R}^d)$, then FP defines an Ergodic Markov Process, whose (unique) invariant distribution is Gibbs:

$$p(x) = \frac{1}{Z} e^{-D^{-1}g(x)} .$$

- The stationary distribution can be used to "renormalize" the Fokker-Plank equation.

- Assume that $p(x,t)$ is the solution of above FP equation, and $\rho(x)$ the associated Gibbs stationary distribution. Defining $h(x,t)$ as $p(x,t) = h(x,t)\rho(x)$, we obtain the backward Kolmogorov equation:

$$\partial_t h = -\langle \nabla g, \nabla h \rangle + D\Delta h .$$

GRADIENT FLOWS AND FOKKER-PLANK

- The previous Fokker-Plank equation is generated by the operator

$$\mathcal{L} = -\nabla g(x) \cdot \nabla + D\Delta \qquad (\partial_t p = \mathcal{L} p)$$

- It turns out that in that case, $\mathcal{L}$ is self-adjoint and non-positive, with null-space consisting only on constant functions. It is self-adjoint if and only if the drift is the gradient of a potential function.
- When the potential function g is convex, one can show that $p(x,t)$ converges to the Gibbs measure exponentially fast in relative entropy (using Poincare inequalities).

FROM SGD TO SGLD

- To summarize, we have constructed another family of SDEs, given by $dX_t = -\nabla g(X_t)dt + \sqrt{2D}dW_t$, whose stationary distribution converges to the Gibbs distribution

$$p(x) = Z^{-1}e^{-D^{-1}g(x)}$$

  - even when g is non-convex.
    - For small enough D>0, such Gibbs distribution concentrates on the global minimizers of g.
- But, can we associate SGD to that diffusion process?
- No: the noise term does not have the right form. It is critical that covariance is positive-definite.

STOCHASTIC GRADIENT LANGEVIN DESCENT

- A possible solution is to add isotropic Gaussian noise to the SGD updates:

$$\theta_{k+1} = \theta_k - \gamma_k(\nabla g(\theta_k) + \epsilon_k) + \sqrt{\gamma_k}\beta_k W_k.$$

ϵ_k: sampling noise, $\mathbb{E}(\epsilon_k|\theta_{k-1}, \ldots, \theta_0) = 0$.

W_k: iid $N(0, I)$ noise.

learning rate γ_k st. $\sum_k \gamma_k^2 < \infty$ but $\sum_k \gamma_k = \infty$.

- By appropriately choosing annealing rate β_k, [Gelfrand & Mitter,'98] show that this algorithm converges to the *continuous simulated annealing* or *Langevin dynamics:*

$$dX_t = -\nabla g(X_t)dt + \eta(t)dW(t)\,, \eta(t) > 0 \in \mathbb{R}, \eta(t) \to 0.$$

SGLD ALGORITHM

- How about the constant learning rate case?
- If $\gamma_k = \gamma$, $\beta_k = \beta$, the former algorithm defines a discrete-time Markov process. What does its limiting distribution converge to as $\gamma, \beta \to 0$?
- Two-step analysis (from [Borkar & Mitter,'99]):
  - As $\gamma \to 0$, the SGLD tracks better and better the diffusion

$$dX_t = -\nabla g(X_t)dt + \beta dW_t$$

  - Its limit converges to the invariant distribution of the Markov process: $p(x) = Z^{-1}e^{-\beta g(x)}$
  - Then set $\beta \to 0$ to concentrate measure on global minima.

SGLD ALGORITHM

- In a ML setup, recall that

$$g(\theta) = \mathbb{E}_{X\sim P}[G(\theta; X)]$$

- The stochastic gradients are obtained via an empirical estimate of g :

$$g_x(\theta) = \frac{1}{n}\sum_{i\leq n} G(\theta; x_i)\,, \theta \in \mathbb{R}^d, x_i \text{ iid} \sim P\,.$$

- We are interested in controlling the generalization error

$$\mathbb{E}g(\hat{\theta}) - \inf_{\theta} g(\theta)$$

  - Expectation with respect to sample x and descent algo.

SGLD IN ML [RAGINSKY ET AL.,'17]

- We consider Stochastic Gradient Langevin Descent on the empirical loss $g_x(\theta)$:

$$\theta_{k+1} = \theta_k - \gamma(\nabla g_x(\theta_k) + \epsilon_k) + \sqrt{2\gamma\beta}W_k$$

γ: learning rate, β: temperature, W_k iid Standard Normal.

- By the previous discussion, this algorithm is the discretization of the Langevin diffusion, described by the SDE

$$dX_t = -\nabla g_x(X_t)dt + \sqrt{2\beta}dW_t\,, t \geq 0\,.$$

- which has as invariant distribution the Gibbs measure

$$\pi_x(d\theta) \propto \exp\{-\beta^{-1}g_x(\theta)\}\,.$$

SGLD IN ML [RAGINSKY ET AL.,'17]

- We need to control several deviation terms. If $\hat{\theta} = \theta_K$ is our parameter estimate after running K SGLD iterations on empirical sample $x = (x_1, \ldots, x_n)$

$$\mathbb{E}g(\hat{\theta}) - \inf_\theta g(\theta) = \mathbb{E}g(\hat{\theta}) - \mathbb{E}g(\hat{\theta}_*) + \mathbb{E}g(\hat{\theta}_*) - \mathbb{E}g_x(\hat{\theta}_*) + \mathbb{E}g_x(\hat{\theta}_*) - \inf_\theta g(\theta).$$

$\theta_* \sim \pi_x(\theta)$: Gibbs sampler

- $\mathbb{E}g(\hat{\theta}) - \mathbb{E}g(\hat{\theta}_*)$: difference in expected population risks of SGLD and Gibbs sampling.
- $\mathbb{E}g(\hat{\theta}_*) - \mathbb{E}g_x(\hat{\theta}_*)$: generalization gap of Gibbs sampling.
- $\mathbb{E}g_x(\hat{\theta}_*) - \inf_\theta g(\theta)$ measures how far are Gibbs samples to global optimum.

SGLD IN ML [RAGINSKY ET AL.,'17]

➤ **Theorem [Raginsky et al. '17, informal]:** The former decomposition terms are bounded as

$$\mathbb{E}g(\hat{\theta}) - \mathbb{E}g(\hat{\theta}_*) \simeq \epsilon \mathrm{Poly}\left(\beta^{-1}, d, \lambda_*^{-1}\right)$$

$$\text{for } K \geq \mathrm{Poly}(\beta^{-1}, d, \lambda_*^{-1})\epsilon^{-4}\,, \text{and } \gamma \leq \frac{\epsilon^4}{\log(\epsilon^{-1})^4}\,,$$

$$\mathbb{E}g(\hat{\theta}_*) - \mathbb{E}g_x(\hat{\theta}_*) \simeq \frac{(\beta^{-1}+d)^2}{\lambda_* n}\,,$$

$$\mathbb{E}g_x(\hat{\theta}_*) - \inf_\theta g(\theta) \simeq \beta d \log(\beta^{-1}+1)\,.$$

➤ Only first term involves SGLD, other two only Gibbs sampler.

➤ λ_* is the spectral gap governing convergence of Markov chain.
(it is independent of n)

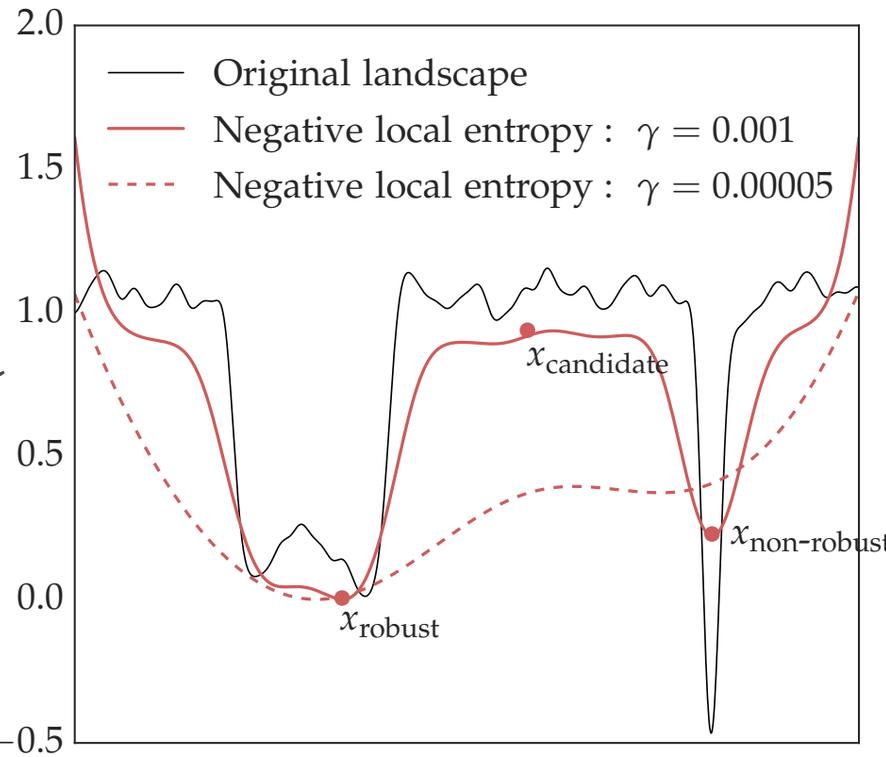
SGLD IN ML [RAGINSKY ET AL.,'17]

- Proof is very technical.
- Main ingredient is to bound the 2-Wasserstein distance between distribution of θ_k from SGLD and $X_{\gamma k}$ from Langevin diffusion.
- Then use logarithmic Sobolev inequality to bound $\mathcal{W}_2$ between distribution of $X_{\gamma k}$ and Gibbs stationary.
- This logarithmic Sobolev inequality is also used to bound the second term: stability of Gibbs distribution to perturbations of the empirical distribution.

OTHER RELATED WORK

- Chaudari et al.'16: "Entropy-SGD: Biasing Gradient Descent into wide valleys"
  - Modifies Gibbs distribution by adding a *proximal* term that favors finding "flat" regions of low-energy:

$$\pi(\theta'|\theta,\beta,\nu) \propto \exp\left(-\beta g(\theta') - \beta\nu\|\theta-\theta'\|^2\right)$$



  - The descent attempts to minimize the resulting *local entropy* of this proximal Gibbs distribution:

$$F(\theta,\beta,\nu) := \log Z(\theta,\beta,\nu)$$

NON-CONVEX OPTIMIZATION

- If optimization gradient-based algorithms converge to global minima in the convex case, what can we expect in non-convex optimization?

- Convergence to local minimisers?
- How many of those are global?
- How much can we be slowed down by saddle points?
- How much noise to add to gradient descent to escape those saddles?

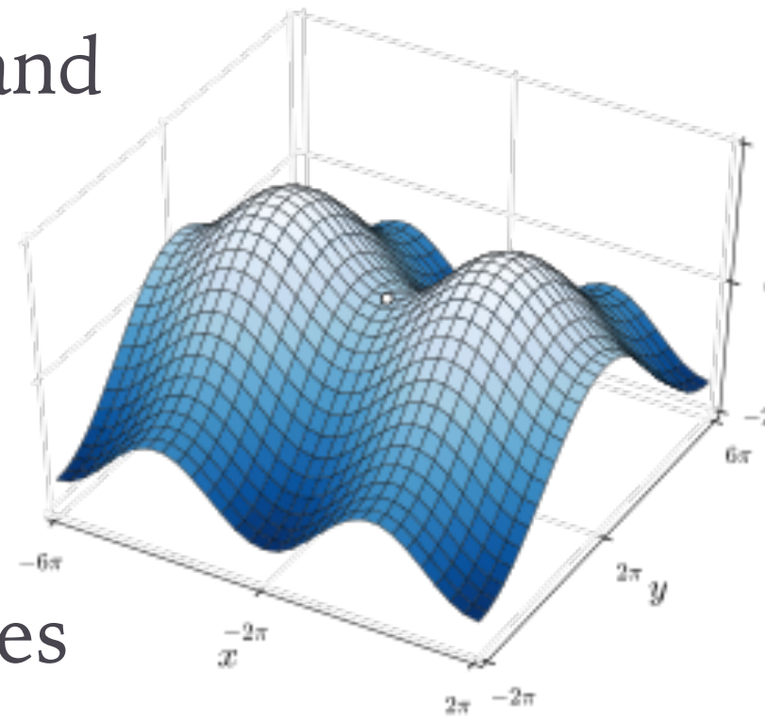
NON-CONVEX OPTIMIZATION

- Vanilla Gradient descent scheme: $\theta_{k+1} = \theta_k - \gamma \nabla g(\theta_k)$
- Its equilibrium points satisfy $\nabla g(\theta_*) = 0$.
- When g is convex, 1st order critical points are global minima.
- However, for general g, stationary points of gradient descent are not necessarily global minima.
- How can they be classified?

INDEX OF CRITICAL POINTS

- Suppose g is in $\mathcal{C}^2$ defined on $\mathbb{R}^d$.
- A *critical point* $\theta^* \in \mathbb{R}^d$ of g is such that $\nabla g(\theta^*) = 0$.
- θ^* is a *strict saddle point* if it is a critical point and its Hessian satisfies

$$\lambda_{\min}(\nabla^2 g(\theta_*)) < 0\,.$$



- A local minima (not necessarily strict) satisfies $\lambda_{\min}(\nabla^2 g(\theta_*)) \geq 0\,.$
- While every critical point is an equilibrium point of gradient descent, which are *stable* equilibria?

CHARACTERIZING STABLE EQUILIBRIA OF GRADIENT DESCENT

- Our intuition is that strict saddles are unstable equilibria, and thus it is unlikely that gradient descent reaches those points.
- We can find worst-case initializations of gradient descent that provably converge to saddle-points [Nesterov,'04].
- On the other hand, we have just seen that adding noise to gradients can be used to escape saddle points.
- But, how likely is it that randomly initialized gradient descent gets stuck in a saddle point?
- We can study the region of attraction associated to each critical point.

STABLE MANIFOLD

- A discrete-time optimization algorithm is a mapping $\varphi : \mathbb{R}^d \to \mathbb{R}^d$
  - ex for gradient descent, $\varphi(\theta) = \theta - \gamma \nabla g(\theta)$.
  - iterate k is thus obtained as $\theta_k = \varphi^k(\theta_0)$.
- Denote by $\mathcal{X}^*$ the set of strict saddle points of g.
- **Definition** (Global Stable Set): The Global Stable Set of the strict saddles is

$$S_\varphi = \{\theta_0; \lim_{k \to \infty} \varphi^k(\theta_0) \in \mathcal{X}^*\} .$$

INTUITION: QUADRATIC CASE

- Consider the non-convex quadratic function

$$g(\theta) = \theta^\top H\theta\,, H = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)\,, \lambda_1, \ldots \lambda_s > 0,\ \lambda_{s+1}, \ldots, \lambda_d < 0\,.$$

- A single critical point, $\theta^* = 0$, which is a strict saddle.
- Gradient descent initialised from θ_0 produces

$$\theta_{k+1} = \varphi(\theta_k) = \sum_{i=1}^{d} (1 - \gamma\lambda_i)^{k+1} \theta_{0,i} e_i\,.$$

- Suppose $\gamma < \dfrac{1}{\max_i |\lambda_i|}$. Then $\|\theta_k\| \to 0$ iff $\theta_0 \in \mathrm{span}\{e_1, \ldots, e_s\}$.
- It follows that if θ_0 is chosen at random, then we will avoid the strict saddle with probability 1.

INTUITION: GENERAL CASE

- In the case of gradient descent, in a neighborhood of a critical point θ_*, the local attractive set is well approximated by the span of eigenvectors corresponding to positive eigenvectors of the Hessian $\nabla^2 g(\theta_*)$.
- This local attractive set has zero measure within a small neighborhood, thus with probability 1 an initialization sufficiently close to θ_* will leave this neighborhood.
- From local stability to global stability?

STABLE MANIFOLD THEOREM

- Given a diffeomorphism φ, we define an unstable fixed point set as

$$\mathcal{A}^*_\varphi = \{\theta;\, \varphi(\theta) = \theta;\, \max_i |\lambda_i(D\varphi(\theta))| > 1\}.$$

- **Theorem [Stable Manifold]:** Let φ be a C^1 mapping from $\mathcal{X} \to \mathcal{X}$ and $\det(D\varphi(\theta)) \neq 0$ for all $\theta \in \mathcal{X}$. Then the set of initial points that converge to an unstable fixed point has measure zero:

$$\mu\left(\{\theta_0 : \lim_k \varphi^k(\theta_0) \in \mathcal{A}^*_\varphi\}\right) = 0\,.$$

- **Corollary:** If $\mathcal{X}^* \subseteq \mathcal{A}^*_\varphi$, then $\mu(S_\varphi) = 0$.

GRADIENT DESCENT AVOIDS STRICT SADDLES

- Assume ∇g is Lipschitz, with $\|\nabla^2 g(\theta)\| \leq L$.
- Fact 1: every strict saddle point of g is an unstable fixed point of gradient descent: $\mathcal{X}^* \subseteq \mathcal{A}^*_\varphi$
- Fact 2: Moreover, if step size satisfies $\gamma < L^{-1}$, then

$$\det(D\varphi(\theta)) \neq 0 \, \forall \theta \, .$$

- Consequence: [Lee et al.'16] Under these assumptions, the stable set of strict saddle points has measure zero.
- This result also holds for proximal point, coordinate descent, mirror descent.

COMMENTS

- This result establishes that gradient descent escapes strict saddles. It does not directly imply that it converges to local minimizers!
- The required additional property is that $\lim_k \theta_k$ exists.
- Two sufficient conditions are discussed in [Lee et al.'16]:
  - Isolated critical points and compact sublevel sets
  - Satisfies the local Lojasiewicz inequality

$$\|\nabla g(\theta)\| \geq m|g(\theta) - g(\theta^*)|^a \, , a < 1$$