L05
Non-convex Optimization: GD + noise converges to second order stationarity

50.579 Optimization for Machine Learning

Ioannis Panageas

ISTD, SUTD

Recap

Theorem (GD avoids strict saddles)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and x^* be a strict saddle point and $\epsilon < 1/L$. For any continuous distribution D, if we sample initialization x_0 from D, GD converges to x^* with probability zero.*

Recap

Theorem (GD avoids strict saddles). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and x^* be a strict saddle point and $\epsilon < 1/L$. For any continuous distribution D, if we sample initialization x_0 from D, GD converges to x^* with probability zero.*

- This is only true in the unconstrained case!

Recap

Theorem (GD avoids strict saddles). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and x^* be a strict saddle point and $\epsilon < 1/L$. For any continuous distribution D, if we sample initialization x_0 from D, GD converges to x^* with probability zero.*

- This is only true in the unconstrained case!

Example (Bad example for constrained). *Consider the following optimization problem:*

$$\min_{x,y} -xye^{-x^2-y^2} + \frac{1}{2}y^2 \ s.t \ x + y \leq 0.$$

- $\nabla f(0,0) = 0.$
- $\nabla^2 f(0,0) = \begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix}$

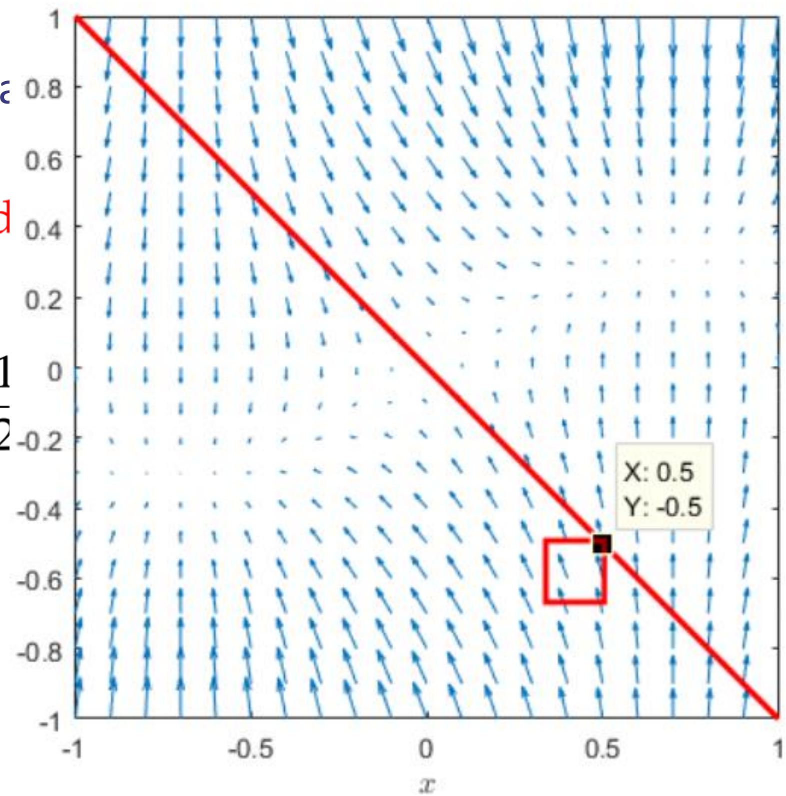
Recap

Theorem (GD avoids strict saddles). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and x^* be a strict saddle point and $\epsilon < 1/L$. For any continuous distribution D, if we sample initialization x_0 from D, GD converges to x^* with probability zero.*

- This is only true in the unconstrained ca

Example (Bad example for constrained *problem:*

$$\min_{x,y} -xye^{-x^2-y^2} + \frac{1}{2}$$



- $\nabla f(0,0) = 0.$
- $\nabla^2 f(0,0) = \begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix}$

Vanishing step-sizes

Theorem (GD avoids strict saddles with vanishing stepsize)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and x^* be a strict saddle point and ϵ_t is of order $\Omega(\frac{1}{t})$ (vanishing). For any continuous distribution D, if we sample initialization x_0 from D, GD converges to x^* with probability zero.*

Vanishing step-sizes

Theorem (GD avoids strict saddles with vanishing stepsize). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and x^* be a strict saddle point and ϵ_t is of order $\Omega(\frac{1}{t})$ (vanishing). For any continuous distribution D, if we sample initialization x_0 from D, GD converges to x^* with probability zero.*

Fact (GD for Quadratic). *Let $f(x) = \frac{1}{2}x^T Ax$. GD boils down to:*

$$x_{t+1} = x_t - \epsilon_t A x_t = (I - \epsilon_t A)x_t.$$

$$\text{Therefore } x_{t+1} = \prod_{z=t}^{0}(I - \epsilon_z A)x_0$$

Vanishing step-sizes

Theorem (GD avoids strict saddles with vanishing stepsize). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and x^* be a strict saddle point and ϵ_t is of order $\Omega(\frac{1}{t})$ (vanishing). For any continuous distribution D, if we sample initialization x_0 from D, GD converges to x^* with probability zero.*

Fact (GD for Quadratic). *Let $f(x) = \frac{1}{2}x^T Ax$. GD boils down to:*

$$x_{t+1} = x_t - \epsilon_t A x_t = (I - \epsilon_t A)x_t.$$

Therefore $x_{t+1} = \prod_{z=t}^{0}(I - \epsilon_z A)x_0$

Since A is symmetric, $A = P^\top \Delta P$ with Δ diagonal matrix, $P^\top P = I$.

Therefore $x_{t+1} = P^\top \prod_{z=t}^{0}(I - \epsilon_z \Delta)Px_0$

Vanishing step-sizes

Therefore $x_{t+1} = P^\top \prod_{z=t}^{0}(I - \epsilon_z \Delta) P x_0$

Observe that $\prod_{z=t}^{0}(I - \epsilon_z \Delta) = \Delta'$, where Δ' is diagonal with entry (i, i)

$$\prod_z (1 - \epsilon_z \lambda_i).$$

Vanishing step-sizes

Therefore $x_{t+1} = P^\top \prod_{z=t}^{0}(I - \epsilon_z \Delta) P x_0$

Observe that $\prod_{z=t}^{0}(I - \epsilon_z \Delta) = \Delta'$, where Δ' is diagonal with entry (i,i)

$$\prod_z (1 - \epsilon_z \lambda_i).$$

Hence the eigenvalues are $e^{\sum \ln(1-\epsilon_z \lambda_i)} \approx e^{-\lambda_i \sum \epsilon_z}$

Assume that $\lambda_i < 0$ As long as $\sum_{z=0}^{\infty} \epsilon_z = \infty$ then for GD to converge to zero, we must have that $P x_0 \perp e_i$.

Definitions

Assumption (Hessian Lipschitz)**.** *We assume that the twice differentiable functions we are deadling with have Hessian ρ-Lipschitz, that is*

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq \rho \left\|x - y\right\|_2 .$$

Definition (Approximate first/second order stationary point)**.** *We provide the following definitions:*

- *A point x^* is an $\epsilon-$first order stationary point (or critical point) of f if $\|\nabla f(x^*)\|_2 \leq \epsilon$.*
- *A point x^* of f is an $\epsilon-$strict saddle point if it is an ϵ-first order stationary point and $\lambda_{\min}(\nabla^2 f(x^*)) < -\sqrt{\rho\epsilon}$*
- *The ϵ-first order points that are not ϵ-strict saddles are called ϵ-second order stationary points.*

Convergence to first order stationarity

Theorem (GD converges to first-order stationarity)**.** *For any $\epsilon > 0$, assume the differentiable function is L-smooth and let $\alpha = \frac{1}{L}$. Moreover, let $f(x^*)$ be the global minimum of f. Then, the gradient descent algorithm in*

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

will visit an ϵ-stationary point at least once in at most $T := \frac{2L(f(x_0)-f(x^*))}{\epsilon^2}$ iterations.

Proof. Recall

$$f(x - \frac{1}{L}\nabla f(x)) - f(x) \le -\frac{1}{2L} \|\nabla f(x)\|_2^2 .$$

Convergence to first order stationarity

Theorem (GD converges to first-order stationarity)**.** *For any $\epsilon > 0$, assume the differentiable function is L-smooth and let $\alpha = \frac{1}{L}$. Moreover, let $f(x^*)$ be the global minimum of f. Then, the gradient descent algorithm in*

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

will visit an ϵ-stationary point at least once in at most $T := \frac{2L(f(x_0) - f(x^*))}{\epsilon^2}$ iterations.

Proof. Recall

$$f(x - \frac{1}{L}\nabla f(x)) - f(x) \le -\frac{1}{2L} \|\nabla f(x)\|_2^2 .$$

Assume that $\|\nabla f(x_t)\|_2 > \epsilon$ for $t = 1, ..., T$. We get that

$$f(x_T) - f(x_{T-1}) + f(x_{T-1}) - f(x_{T-2}) + \ldots + f(x_1) - f(x_0) < -\frac{\epsilon^2 T}{2L}.$$

Convergence to first order stationarity

Theorem (GD converges to first-order stationarity)**.** *For any $\epsilon > 0$, assume the differentiable function is L-smooth and let $\alpha = \frac{1}{L}$. Moreover, let $f(x^*)$ be the global minimum of f. Then, the gradient descent algorithm in*

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

will visit an ϵ-stationary point at least once in at most $T := \frac{2L(f(x_0)-f(x^*))}{\epsilon^2}$ iterations.

Proof. Recall

$$f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L} \|\nabla f(x)\|_2^2 .$$

Therefore $f(x^*) - f(x_0) \leq f(x_T) - f(x_0) < -\frac{\epsilon^2 T}{2L} = f(x^*) - f(x_0).$

Convergence to first order stationarity

Theorem (GD converges to first-order stationarity)**.** *For any $\epsilon > 0$, assume the differentiable function is L-smooth and let $\alpha = \frac{1}{L}$. Moreover, let $f(x^*)$ be the global minimum of f. Then, the gradient descent algorithm in*

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

will visit an ϵ-stationary point at least once in at most $T := \frac{2L(f(x_0) - f(x^*))}{\epsilon^2}$ iterations.

Proof. Recall

$$f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L} \|\nabla f(x)\|_2^2 .$$

Therefore $f(x^*) - f(x_0) \leq f(x_T) - f(x_0) < -\frac{\epsilon^2 T}{2L} = f(x^*) - f(x_0)$.

Contradiction!

Perturbed Gradient Descent

Definition (Perturbed Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. The Perturbed Gradient Descent is defined as follows:*

1. Initialization x^0, stepsize η, perturbation radius r.
2. **For** t=1 ... T **do**
3. $\quad x_{t+1} = x_t - \eta(\nabla f(x_t) + \xi_t)$ with $\xi_t \sim \mathcal{N}(0, (r^2/d)I)$
4. **End For**

Theorem (PGD converges to second-order stationarity). *Let f be a twice differentiable L-smooth function with Hessian ρ-Lipschitz. For any $\epsilon, \delta > 0$, set $\eta = \Theta(\frac{1}{L})$, $r = \Theta\left(\frac{\epsilon}{\log^4 d/(\delta\epsilon)}\right)$. PGD will visit an ϵ-second-order stationary point at least once with probability at least $1-\delta$ in at most $T = O\left(\frac{L(f(x_0)-f(x^*))}{\epsilon^2}\log^4 \frac{d}{\rho\epsilon\delta}\right)$ iterations.*

Analysis of Perturbed Gradient Descent

- High level proof strategy:

1) When the current iterate is not an ϵ-second order stationary point,
it must either (a) have a large gradient or (b) have a strictly negative eigenvalue the Hessian.

2) We can show in both cases that yield a significant decrease in function value
in a controlled number of iterations.

3) Since the decrease cannot be more that $\mathrm{f}(\mathrm{x}_0) - f(x^*)$ (global minimum is bounded)
we can reach contradiction.

Analysis of Perturbed Gradient Descent

Lemma (Descent Lemma). *Assume f is twice differentiable L-smooth and $\eta = \frac{1}{L}$. Then it holds with probability $1 - \delta$*

$$f(x_{t+1}) - f(x_t) \leq -\frac{\|\nabla f(x_t)\|^2}{2L} + O\left(r^4/d^2 \log \frac{1}{\delta}\right).$$

Proof.

$$f(x_{t+1}) - f(x_t) \leq \nabla f(x_t)^\top (x_{t+1} - x_t) + \tfrac{L}{2} \|x_{t+1} - x_t\|_2^2 \text{ L-smooth,}$$

Analysis of Perturbed Gradient Descent

Lemma (Descent Lemma)**.** *Assume f is twice differentiable L-smooth and $\eta = \frac{1}{L}$. Then it holds with probability $1-\delta$*

$$f(x_{t+1}) - f(x_t) \leq -\frac{\|\nabla f(x_t)\|^2}{2L} + O\left(r^4/d^2 \log\frac{1}{\delta}\right).$$

Proof.

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \nabla f(x_t)^\top (x_{t+1} - x_t) + \tfrac{L}{2}\|x_{t+1} - x_t\|_2^2 \ \text{L-smooth}, \\ &= -\tfrac{1}{L}\nabla f(x_t)^\top \nabla f(x_t) - \tfrac{1}{L}\xi_t^\top \nabla f(x_t) + \tfrac{L}{2}\tfrac{1}{L^2}\|\nabla f(x) + \xi_t\|_2^2, \end{aligned}$$

Analysis of Perturbed Gradient Descent

Lemma (Descent Lemma). *Assume f is twice differentiable L-smooth and $\eta = \frac{1}{L}$. Then it holds with probability $1-\delta$*

$$f(x_{t+1}) - f(x_t) \leq -\frac{\|\nabla f(x_t)\|^2}{2L} + O\left(r^4/d^2 \log\frac{1}{\delta}\right).$$

Proof.

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \nabla f(x_t)^\top (x_{t+1} - x_t) + \tfrac{L}{2}\|x_{t+1} - x_t\|_2^2 \text{ L-smooth}, \\ &= -\tfrac{1}{L}\nabla f(x_t)^\top \nabla f(x_t) - \tfrac{1}{L}\xi_t^\top \nabla f(x_t) + \tfrac{L}{2}\tfrac{1}{L^2}\|\nabla f(x) + \xi_t\|_2^2, \\ &\leq -\tfrac{1}{2L}\|\nabla f(x)\|_2^2 + \tfrac{1}{2L}\|\xi_t\|_2^2. \end{aligned}$$

Analysis of Perturbed Gradient Descent

Lemma (Descent Lemma). *Assume f is twice differentiable L-smooth and $\eta = \frac{1}{L}$. Then it holds with probability $1-\delta$*

$$f(x_{t+1}) - f(x_t) \le -\frac{\|\nabla f(x_t)\|^2}{2L} + O\left(r^4/d^2 \log\frac{1}{\delta}\right).$$

Proof.

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\le \nabla f(x_t)^\top (x_{t+1} - x_t) + \tfrac{L}{2}\|x_{t+1} - x_t\|_2^2 \text{ L-smooth,} \\ &= -\tfrac{1}{L}\nabla f(x_t)^\top \nabla f(x_t) - \tfrac{1}{L}\xi_t^\top \nabla f(x_t) + \tfrac{L}{2}\tfrac{1}{L^2}\|\nabla f(x) + \xi_t\|_2^2, \\ &\le -\tfrac{1}{2L}\|\nabla f(x)\|_2^2 + \tfrac{1}{2L}\|\xi_t\|_2^2. \end{aligned}$$

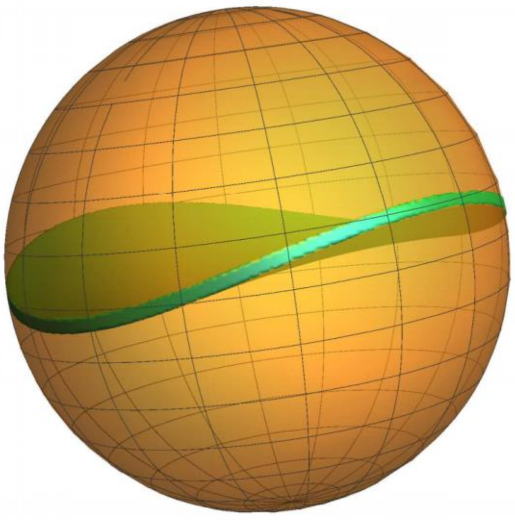
This is of order $\Theta(\epsilon^2)$ if we are in case (a).

Analysis of Perturbed Gradient Descent

Lemma (Escaping saddle points). *Assume f is twice differentiable L-smooth and has hessian ρ-Lipschitz. Moreover assume that $\|\nabla f(x_0)\|_2 \leq \epsilon$ and also $\lambda_{min}(\nabla^2 f(x_0)) \leq -\sqrt{\rho\epsilon}$. Assume we run PGD from x_0, then*

$$\Pr[f(x_t) - f(x_0) \leq -\frac{t'}{2}] \geq 1 - \frac{L\sqrt{d}}{\sqrt{\rho\epsilon}} e^{-\Theta(\log^4 \frac{d}{\rho\epsilon})},$$

for $t = \frac{L}{\sqrt{\rho\epsilon}}\Theta(\log^4 \frac{d}{\rho\epsilon})$ and $t' = \frac{\epsilon^2}{\sqrt{\rho\epsilon}}\Theta(\log^4 \frac{d}{\rho\epsilon})$.



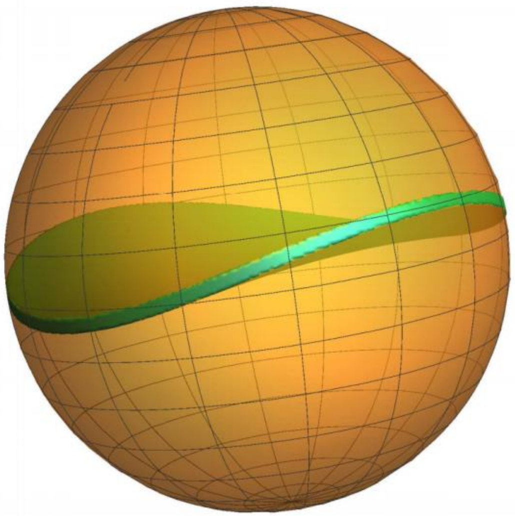
Analysis of Perturbed Gradient Descent

Lemma (Escaping saddle points). *Assume f is twice differentiable L-smooth and has hessian ρ-Lipschitz. Moreover assume that $\|\nabla f(x_0)\|_2 \leq \epsilon$ and also $\lambda_{min}(\nabla^2 f(x_0)) \leq -\sqrt{\rho\epsilon}$. Assume we run PGD from x_0, then*

$$\Pr[f(x_t) - f(x_0) \leq -\frac{t'}{2}] \geq 1 - \frac{L\sqrt{d}}{\sqrt{\rho\epsilon}} e^{-\Theta(\log^4 \frac{d}{\rho\epsilon})},$$

for $t = \frac{L}{\sqrt{\rho\epsilon}}\Theta(\log^4 \frac{d}{\rho\epsilon})$ and $t' = \frac{\epsilon^2}{\sqrt{\rho\epsilon}}\Theta(\log^4 \frac{d}{\rho\epsilon})$.

Since $f(x^*) - f(x_0)$ is bounded and t is $\Theta(t'\epsilon^2)$, after $\Theta(\frac{f(x^*)-f(x_0)}{\epsilon^2})$ we reach a second order stationary point (contradiction otherwise).



Conclusion

- Introduction to Non-convex Optimization.
  - Perturbed Gradient Descent avoids strict saddles!
  - Same is true for Perturbed SGD.
- Next lecture we will talk about more about accelerated methods.
- Week 8 we are going to talk about min-max optimization (GANs).