Gradient Descent Method

Course: SDS 384
Instructor: Nhat Ho

Fall 2022

- Apart from foundational aspects of gradient descent method, our lecture will also involve:
    - Concentration inequalities with a sequence of random variables (You can read more about this topic from Section 2 to Section 5 in [1])
    - Research questions that you can use for your final project or your research topic
- We will focus on the trade-off between statistical complexity and computational efficiency of gradient descent method in machine learning models
    - It yields useful insight into the practice of gradient descent method
    - It leads to a design of optimal optimization algorithms for certain machine learning models

Gradient Descent: Motivation

- Gradient descent is perhaps the simplest optimization algorithm
- It is simple to implement and has low computational complexity, which fits to large-scale machine learning problems
- Asymptotic behaviors of gradient descent had been understood quite well
- **Research problems**: Understanding the **non-asymptotic behaviors** of gradient descent and its variants has still remained an active important research area
    - How many data samples are needed for gradient descent to reach a certain neighborhood around the true value?
    - For adaptive gradient methods, like Adam/ Adagrad/ Polyak average step size, the trade-off between their statistical guarantee and their computational complexity remained poorly understood
    - How to develop uncertainty quantification for (adaptive) gradient descent iterates, such as confidence intervals, etc.?

Example: Generalized Linear Model

Sample size

- Assume that we have a generalized linear model:

$$Y_i = g(X_i^\top \theta^*) + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n,$$

where $Y_1, \ldots, Y_n \in \mathbb{R}$ are response variables;

$X_1, \ldots, X_n \in \mathbb{R}^d$ are covariates;

$\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. (independently identical distributed) random noises from $\mathcal{N}(0,1)$;

g: a given link function (e.g, $g(x) = x$ —linear regression model;

$g(x) = x^2$ — phase retrieval problem)

Goal: Estimate true but unknown parameter $\theta^* \in \mathbb{R}^d$

Example: Generalized Linear Model

- Least-square loss:

$$\min_{\theta \in \mathbb{R}^d} \mathscr{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - g(X_i^\top \theta)\}^2$$

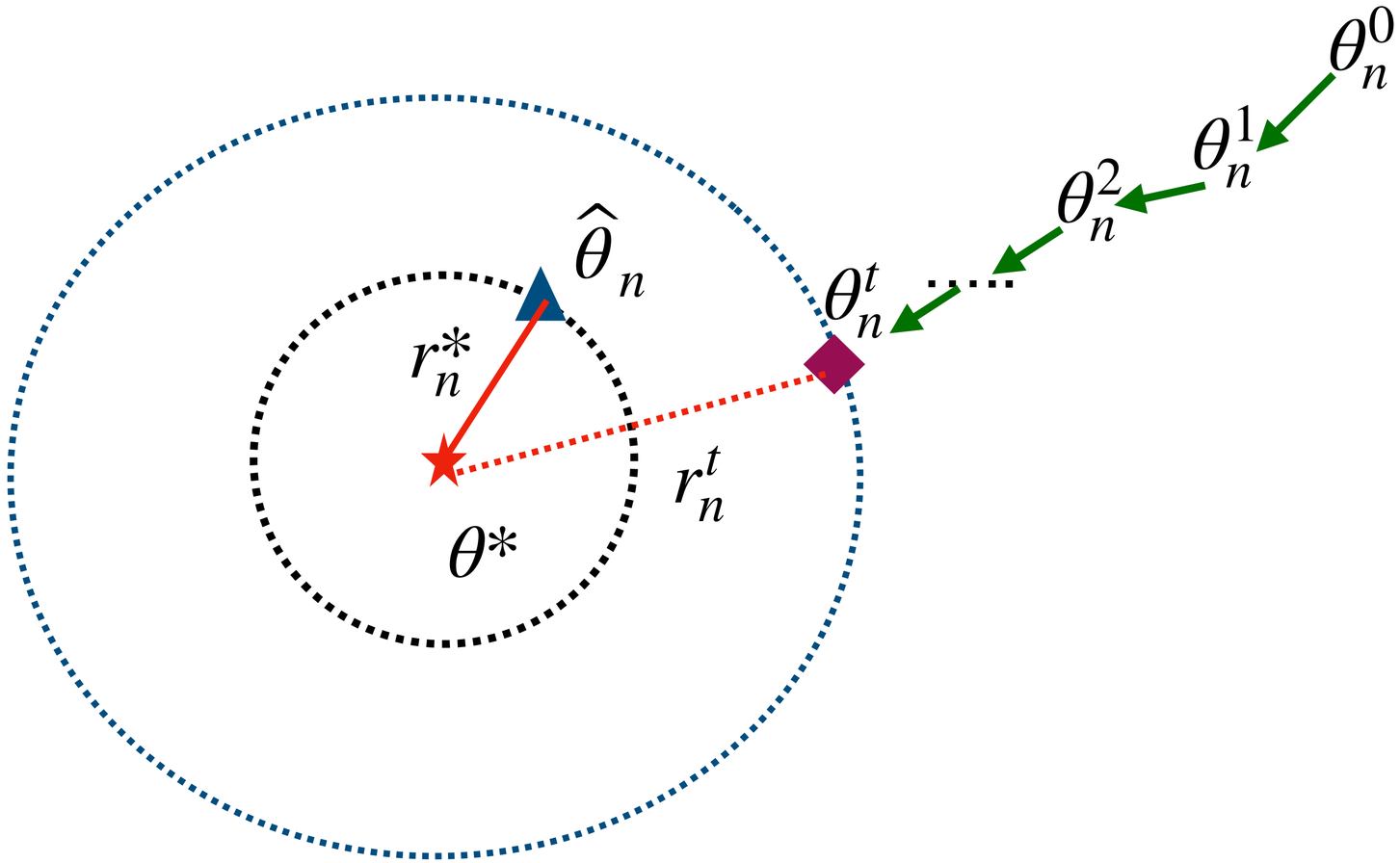
- In general, for non-linear link function g, such as $g(x) = x^p$ for $p > 1$, we do not have closed-form expression for the optimal solution
- We use gradient descent algorithm to approximate the optimal solution
- Denote θ_n^t: iterate of gradient descent at step t,

  $\widehat{\theta}_n$: optimal solution of the least-square loss

Example: Generalized Linear Model

r_n^*: Radius of convergence

r_n^t: Current radius from step t



Research Questions: (Q.1) For a fixed t, can we develop (tight) bounds for r_n^t?

(Q.2) What is the lower bound for t such that $r_n^t \leq C r_n^*$ for some universal constant C

(Q.3) What is a good confidence interval for θ_n^t for fixed t?

Example: Generalized Linear Model

- These open questions require the following deep understandings:
    - **Model perspective**: The optimization landscape of generalized linear model or machine learning models in general
    - **Optimization perspective**: The dynamics of gradient descent algorithm and its variants
    - **Statistical perspective**: The concentration behaviors of the least-square loss and its derivatives, i.e., for fixed sample size, how close the least-square loss and its derivatives to their expectations?

- We will first focus on the optimization perspective of gradient descent method

Unconstrained Optimization: Global Strong Convexity and Smoothness

Unconstrained Optimization

- We first consider general unconstrained optimization problems
- Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is an objective function that is differentiable
- The unconstrained optimization problem is given by:

$$\min_{\theta \in \mathbb{R}^d} f(\theta)$$

- We use gradient descent method to approximate the optimal solution θ^* of that problem

Gradient Descent Method

- Assume that we start with an initialization θ^0
- We would like to build an iterative scheme $\theta^{t+1} = F(\theta^t)$ such that

$$f(\theta^{t+1}) < f(\theta^t)$$

for any $t \geq 0$

- **Gradient descent method**:

$$\theta^{t+1} = \theta^t - \eta_t \nabla f(\theta^t),$$

where $\eta_t > 0$: step size/ learning rate (may also be adaptive with t)

Gradient Descent Method

- Gradient descent finds **steepest descent** at the current iteration θ^t
- Descent direction d at θ^t:

$$d^\top \nabla f(\theta^t) < 0$$

- Simple application of Cauchy-Schwarz:

$$\min_{\|d\| \leq 1} d^\top \nabla f(\theta^t) = -\|\nabla f(\theta^t)\|$$

where the equality holds when $d = -\dfrac{\nabla f(\theta^t)}{\|\nabla f(\theta^t)\|}$

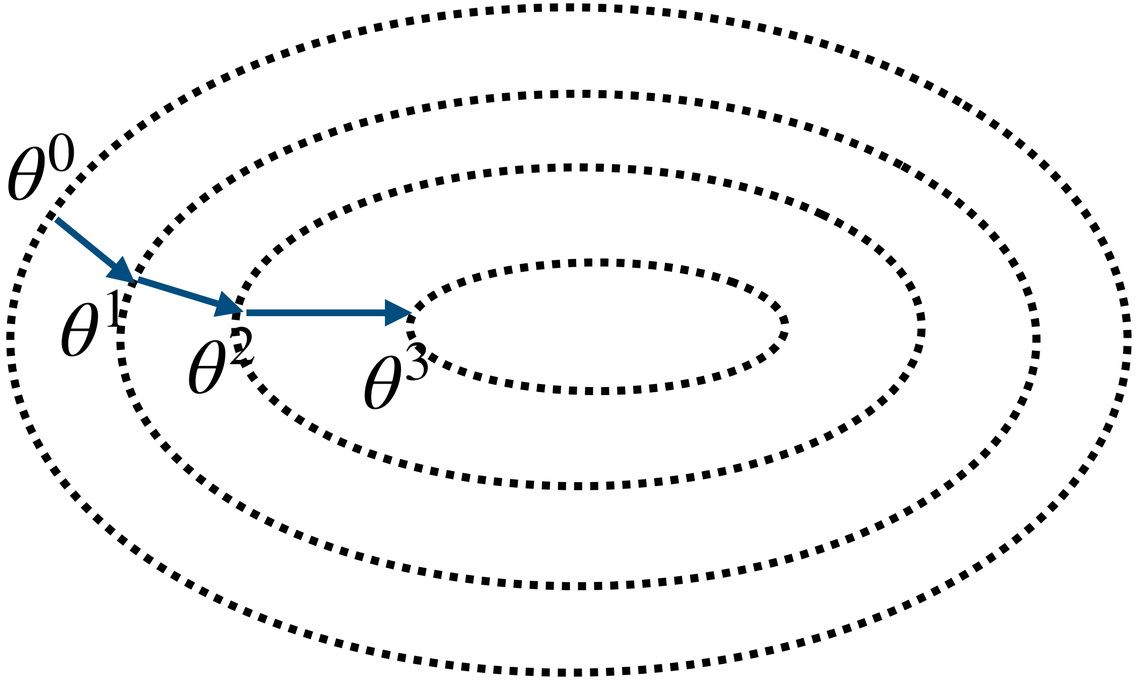


Illustration of gradient descent method

Strong Convexity

- If f is second order differentiable, i.e., second order derivative exists, strong convexity is equivalent to

$$\lambda_{\min}(\nabla^2 f(\theta)) \geq \mu > 0,$$

for all θ where $\lambda_{\min}(A)$: smallest eigenvalue of matrix A

- We call f to be μ-**strongly convex** function
- That notion also implies that

$$f(\theta) \geq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{\mu}{2}||\theta - \theta'||^2,$$

for all $\theta, \theta' \in \mathbb{R}^d$

Strongly Convexity

- Examples of strongly convex function:

$$f(\theta) = ||\theta||^2$$

- Indeed, we have $\nabla^2 f(\theta) = 2 \cdot I_d$, which has smallest eigenvalue to be 2
- Therefore, function f is 2-strongly convex

- Note that, when $f(\theta) = ||\theta||^{2p}$ for $p \in \mathbb{N}$ and $p \geq 2$, the function f is no longer strongly convex (Check that!)

This kind of behavior appears in several machine learning models, such as generalized linear models, mixture models, matrix completion/ sensing, deep neural networks, etc.

Smoothness

- The function f is L-smooth if f is second order differentiable and

$$\lambda_{\max}(\nabla^2 f(\theta)) \leq L,$$

for all θ where $\lambda_{\max}(A)$: largest eigenvalue of matrix A

- That notion also implies that

$$f(\theta) \leq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{L}{2}\|\theta - \theta'\|^2,$$

for all $\theta, \theta' \in \mathbb{R}^d$

- If $f(\theta) = \|\theta\|^2$, then
    - $\lambda_{\max}(\nabla^2 f(\theta)) = 2$
    - The function f is 2-smooth

Convergence Rate: Strong Convexity and Smoothness

Theorem 1: Assume that the function f is μ-strongly convex and L-smooth. As long as $\eta_t = \eta \leq \frac{1}{L}$, we obtain that

$$\|\theta^t - \theta^*\| \leq \left(1 - \eta \cdot \mu\right)^{\frac{t}{2}} \|\theta^0 - \theta^*\|$$

- It shows that θ^t converges geometrically fast to θ^*
- When $\eta = \frac{1}{L}$, the contraction coefficient becomes $\sqrt{1 - \frac{\mu}{L}} = \sqrt{1 - \frac{1}{\kappa}}$, where κ is a condition number
- We can improve the upper bound of η to $\frac{2}{L + \mu}$ (Not the focus of the class)

Convergence Rate: Strong Convexity and Smoothness

Lemma 1: If f is μ-strongly convex and L-smooth function, then

$$2\mu(f(\theta) - f(\theta^*)) \leq \|\nabla f(\theta)\|^2 \leq 2L(f(\theta) - f(\theta^*))$$

Proof of Lemma 1: From the strong convexity, we have

$$\begin{aligned} f(\theta) - f(\theta^*) &\leq \langle \nabla f(\theta), \theta - \theta^* \rangle - \frac{\mu}{2}\|\theta - \theta^*\|^2 \\ &= -\frac{1}{2}\|\sqrt{\mu}(\theta - \theta^*) - \frac{1}{\sqrt{\mu}}\nabla f(\theta)\|^2 + \frac{1}{2\mu}\|\nabla f(\theta)\|^2 \\ &\leq \frac{1}{2\mu}\|\nabla f(\theta)\|^2 \end{aligned}$$

- Similar argument for the upper bound

Convergence Rate: Strong convexity and Smoothness

Proof of Theorem 1:

- $\|\theta^{t+1} - \theta^*\|^2 = \|\theta^t - \theta^* - \eta \nabla f(\theta^t)\|^2$

$$= \|\theta^t - \theta^*\|^2 - 2\eta\langle \nabla f(\theta^t), \theta^t - \theta^*\rangle + \eta^2 \|\nabla f(\theta^t)\|^2 \qquad \textbf{(1)}$$

- $\mu-$Strong convexity indicates

$$f(\theta^*) \geq f(\theta^t) + \langle \nabla f(\theta^t), \theta^* - \theta^t\rangle + \frac{\mu}{2}\|\theta^t - \theta^*\|^2 \qquad \textbf{(2)}$$

- Equations (1) and (2) lead to

$$\|\theta^{t+1} - \theta^*\|^2 \leq (1 - \eta \cdot \mu)\|\theta^t - \theta^*\|^2 + 2\eta(f(\theta^*) - f(\theta^t)) + \eta^2 \|\nabla f(\theta^t)\|^2 \qquad \textbf{(3)}$$

Convergence Rate: Strong convexity and Smoothness

- Use Lemma 1 to equation (3):

$$\|\theta^{t+1} - \theta^*\|^2 \leq (1 - \eta \cdot \mu)\|\theta^t - \theta^*\|^2 + (2L\eta^2 - 2\eta)(f(\theta^t) - f(\theta^*))$$

- As $\eta \leq \frac{1}{L}$, we have $2L\eta^2 \leq 2\eta$
- Therefore, we find that:

$$\|\theta^{t+1} - \theta^*\|^2 \leq (1 - \eta \cdot \mu)\|\theta^t - \theta^*\|^2$$

- Repeat this argument, we obtain the conclusion of Theorem 1

Convergence Rate: Strong Convexity and Smoothness

Corollary 1: Assume that the function f is μ-strongly convex and L-smooth. As long as $\eta_t = \eta \leq \frac{1}{L}$, we obtain that

$$f(\theta^t) - f(\theta^*) \leq \frac{L}{2}\left(1 - \eta \cdot \mu\right)^t \|\theta^0 - \theta^*\|^2$$

Convergence Rate: Strong Convexity and Smoothness

- We now consider the generalized linear model when $g(x) = x$, i.e., linear regression
- The least-square loss is:

$$\min_{\theta \in \mathbb{R}^d} \mathscr{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - X_i^\top \theta\}^2$$

- We can check that

$$\nabla \mathscr{L}_n(\theta) = -\frac{2}{n} \sum_{i=1}^{n} X_i (Y_i - X_i^\top \theta);$$

$$\nabla^2 \mathscr{L}_n(\theta) = \frac{2}{n} \sum_{i=1}^{n} X_i X_i^\top$$

Convergence Rate: Strong Convexity and Smoothness

- To simplify the understanding, consider one dimensional problem, i.e, $d = 1$
- Assume that $X_i \sim \mathcal{N}(0,1)$
- $$\nabla^2 \mathscr{L}_n(\theta) = \frac{2}{n} \sum_{i=1}^{n} X_i^2$$
- From the concentration inequality of chi-square distribution (see Example 2.11 in [1]),

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \mathbb{E}(X^2) \right| \geq t \right) \leq 2 \exp(-n \cdot t^2/8),$$

where the outer expectation is taken with respect to $X \sim \mathcal{N}(0,1)$

Convergence Rate: Strong Convexity and Smoothness

- We have $\mathbb{E}(X^2) = 1$ as $X \sim \mathcal{N}(0,1)$
- It suggests that with probability $1 - \delta$ for some $\delta > 0$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i^2 - 1 \right| \leq \frac{C \cdot \sqrt{\log(1/\delta)}}{\sqrt{n}},$$

where C is some universal constant

- As $\nabla^2 \mathscr{L}_n(\theta) = \frac{2}{n} \sum_{i=1}^{n} X_i^2$, it demonstrates that with probability $1 - \delta$:

$$\mu =: 2 - \frac{C \cdot \sqrt{\log(1/\delta)}}{\sqrt{n}} \leq \nabla^2 \mathscr{L}_n(\theta) \leq 2 + \frac{C \cdot \sqrt{\log(1/\delta)}}{\sqrt{n}} := L$$

➡ The function $\mathscr{L}_n$ is L-smooth and μ-strongly convex

Convergence Rate: Strong Convexity and Smoothness

- Recall the least-square loss $\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \cdot \theta)^2$
- Denote by $\widehat{\theta}_n$ the optimal solution of least-square loss
- From Theorem 1, if we call θ_n^t the gradient descent iterate at step t, then with high probability

$$\|\theta_n^t - \widehat{\theta}_n\| \leq (1 - \eta \cdot \mu)^{t/2} \|\theta_n^0 - \widehat{\theta}_n\|$$

as long as $\eta \leq \frac{1}{L}$

Convergence Rate: Strong Convexity and Smoothness

- Recall that, we would like to estimate θ^*, a true but unknown parameter, of the true model: $Y_i = X_i \cdot \theta^* + \varepsilon_i$
- **(Q.1)**: For a fixed t, what is an upper bounds for $\|\theta_n^t - \theta^*\|$?
- Our result with gradient descent sequence $\{\theta_n^t\}_{t \geq 0}$ shows that

$$\|\theta_n^t - \theta^*\| \leq \|\theta_n^t - \widehat{\theta}_n\| + \|\widehat{\theta}_n - \theta^*\| \quad \text{(Triangle inequality)}$$

$$\leq (1 - \eta \cdot \mu)^{t/2} \|\theta_n^0 - \widehat{\theta}_n\| + \|\widehat{\theta}_n - \theta^*\| \qquad \textbf{(4)}$$

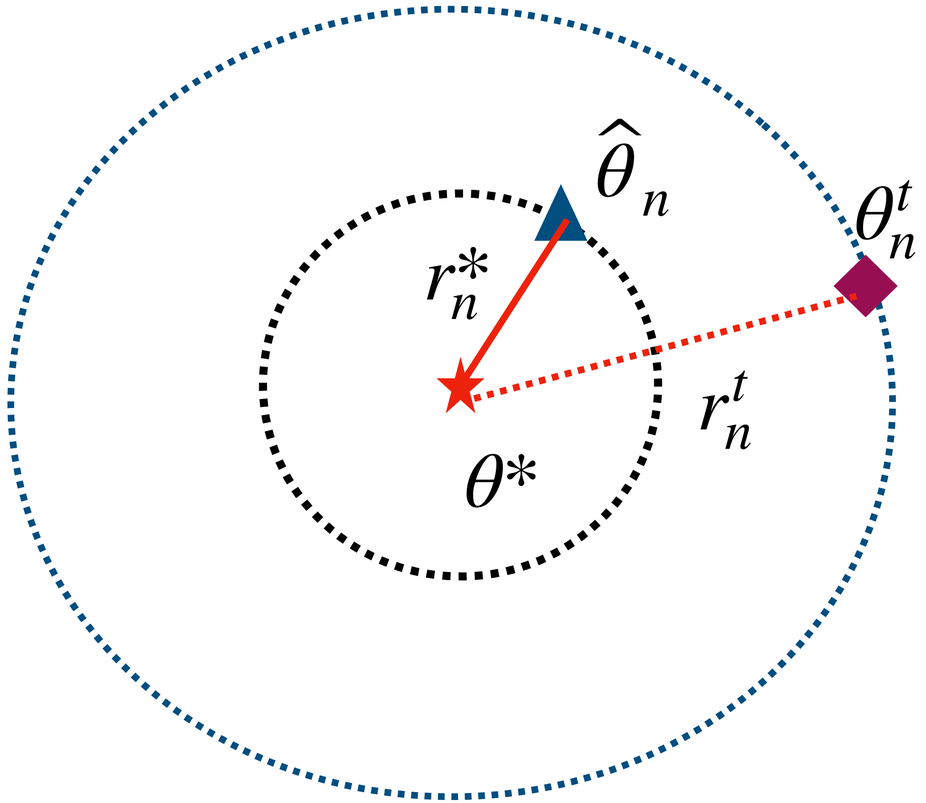
Optimization error — Statistical error

- It answers question (Q.1) earlier in first slides

Convergence Rate: Strong Convexity and Smoothness

- **(Q.2):** The number of iterations t such that θ_n^t reaches the radius of convergence:

$$\|\theta_n^t - \theta^*\| \leq C' \|\widehat{\theta}_n - \theta^*\|$$



Convergence Rate: Strong Convexity and Smoothness

- Recall from equation (4) that:

$$\|\theta_n^t - \theta^*\| \leq (1 - \eta \cdot \mu)^{t/2} \|\theta_n^0 - \widehat{\theta}_n\| + \|\widehat{\theta}_n - \theta^*\|$$

- We choose t such that $(1 - \eta \cdot \mu)^{t/2} \|\theta_n^0 - \widehat{\theta}_n\| \leq \|\widehat{\theta}_n - \theta^*\|$, i.e.,

$$t \leq \frac{2 \log\left(\|\widehat{\theta}_n - \theta^*\| / \overline{C} \right)}{\log(1 - \eta \cdot \mu)} \approx \log(n) \,,$$

where $\overline{C}$ is such that $\|\theta_n^0 - \widehat{\theta}_n\| \leq \overline{C}$

- Therefore, we need $t \geq c \cdot \log n$ for some constant c to answer question (Q.2)

Convergence Rate: Strong Convexity and Smoothness

- **Takeaway messages**: Under the strongly convex and smooth settings of the loss function, we can have a good understanding of:
    - The range of gradient descent updates at fixed number of iterates
    - The sufficient number of iterations for the gradient descent updates to reach the radius of convergence

- **Remaining problems:**
    - When the loss function is not strongly convex or smooth, analyzing the updates of gradient descent becomes more challenging

Local Strong Convexity and Local Smoothness

Local Smoothness

- We now consider the generalized linear model when $g(x) = x^2$, i.e., quadratic regression or phase retrieval problem
- The phase retrieval least-square loss function in one dimension is:

$$\min_{\theta \in \mathbb{R}} \overline{\mathscr{L}}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - (X_i \cdot \theta)^2\}^2$$

$$= \left(\frac{1}{n} \sum_{i=1}^{n} X_i^4 \right) \theta^4 - \left(\frac{2}{n} Y_i X_i^2 \right) \theta^2 + \frac{1}{n} \sum_{i=1}^{n} Y_i^2$$

- We first assume that $\theta^* \neq 0$ (The case $\theta^* = 0$ will be considered later)
- We will show that the loss function $\overline{\mathscr{L}}_n$ is not strongly convex (It is indeed non-convex problem)

Local Smoothness

- Direct calculation shows that

$$\nabla \overline{\mathscr{L}}_n(\theta) = \left(\frac{4}{n} \sum_{i=1}^{n} X_i^4 \right) \theta^3 - \frac{4}{n} (\sum_{i=1}^{n} Y_i X_i^2) \theta,$$

$$\nabla^2 \overline{\mathscr{L}}_n(\theta) = \left(\frac{12}{n} \sum_{i=1}^{n} X_i^4 \right) \theta^2 - \frac{4}{n} (\sum_{i=1}^{n} Y_i X_i^2)$$

- Recall that, $Y_i = (X_i \cdot \theta^*)^2 + \varepsilon_i$ where $X_i \sim \mathcal{N}(0,1)$ and $\theta^* \neq 0$
- By solving $\nabla \overline{\mathscr{L}}_n(\theta) = 0$, we obtain with high probability that $\theta = 0$ is local maxima and

$\theta = \pm \sqrt{\frac{\sum_{i=1}^{n} Y_i X_i^2}{\sum_{i=1}^{n} X_i^4}}$ are two global minima

Local Smoothness

- Both $\frac{12}{n}\sum_{i=1}^{n} X_i^4$ and $4(\sum_{i=1}^{n} Y_i X_i^2)$ are bounded with high probability (We ignore the rigorousness here)
- It demonstrates that when $\theta \to \infty$, $\nabla^2 \overline{\mathscr{L}}_n(\theta) \to \infty$ with high probability
- Hence, the loss function $\overline{\mathscr{L}}_n$ is not **globally smooth** (Recall that, we need step size $\eta \leq 1/L$. Therefore, as $L \to \infty$, the upper bound for step size goes to 0, which is bad!)

Local Smoothness

- $\nabla^2 \overline{\mathscr{L}}_n(\theta) = \left(\frac{12}{n} \sum_{i=1}^{n} X_i^4 \right) \theta^2 - 4(\sum_{i=1}^{n} Y_i X_i^2)$
- If we indeed consider $\theta \in \mathbb{B}(\hat{\theta}_n, r_0) = \left\{ \theta' \in \mathbb{R} : \|\theta' - \hat{\theta}_n\| \leq r_0 \right\}$, then we can find L_1 depending on r_0 such that with high probability

$$\nabla^2 \overline{\mathscr{L}}_n(\theta) \leq L_1,$$

for all $\theta \in \mathbb{B}(\hat{\theta}_n, r_0)$ where $\hat{\theta}_n$ is an optimal solution of $\overline{\mathscr{L}}_n$

- It demonstrates that the function is **locally smooth** around $\hat{\theta}_n$

Local Strong Convexity

- Recall that, $\nabla^2 \overline{\mathscr{L}}_n(\theta) = \left(\frac{12}{n}\sum_{i=1}^{n} X_i^4\right)\theta^2 - \frac{4}{n}(\sum_{i=1}^{n} Y_i X_i^2)$
- We know that $\hat{\theta}_n = \pm\sqrt{\frac{\sum_{i=1}^{n} Y_i X_i^2}{\sum_{i=1}^{n} X_i^4}}$
- The Hessian is zero when $\bar{\theta}_n = \pm\sqrt{\frac{\sum_{i=1}^{n} Y_i X_i^2}{3\sum_{i=1}^{n} X_i^4}}$
- It shows that we not always have $\nabla^2 \overline{\mathscr{L}}_n(\theta)$ bounded away from 0 when $\theta \in \mathbb{B}(\hat{\theta}_n, r)$ where $r \geq \|\bar{\theta}_n - \hat{\theta}_n\|$ (here we assume $\bar{\theta}_n$, $\hat{\theta}_n$ have the same sign)

Local Strong Convexity

- This example shows that as long as $r = \|\bar{\theta}_n - \widehat{\theta}_n\| - \epsilon$ for some small $\epsilon > 0$, the function $\overline{\mathscr{L}}_n$ is **locally strongly convex** around $\widehat{\theta}_n$

- Combining with the local smoothness argument earlier, the ball $\mathbb{B}(\widehat{\theta}_n, \|\widehat{\theta}_n - \bar{\theta}_n\| - \epsilon)$ is sufficient to guarantee that the function $\overline{\mathscr{L}}_n$ is locally strongly convex and smooth around $\widehat{\theta}_n$

Local Strong Convexity

- We now consider another popular example of local strong convexity: logistic regression

- For logistic regression, we assume that $Y_i \in \{-1, 1\}$ and

$$\mathbb{P}(Y_i = 1 \mid X_i) = \frac{\exp(X_i^\top \theta^*)}{1 + \exp(X_i^\top \theta^*)}$$

where θ^* is unknown true parameter

- We estimate the true parameter θ^* via maximum likelihood estimation

Local Strong Convexity

- It is equivalent to:

$$\max_{\theta \in \mathbb{R}^d} \mathscr{T}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-Y_i X_i^\top \theta))$$

- Some algebra shows that

$$\nabla^2 \mathscr{T}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\exp(-Y_i X_i^\top \theta)}{(1 + \exp(-Y_i X_i^\top \theta))^2} X_i X_i^\top$$

- When $\|\theta\| \to \infty$, $\lambda_{\min}(\nabla^2 \mathscr{T}_n(\theta)) \to 0$

- The function $\mathscr{T}_n$ is not globally strongly convex (In fact, it is locally strongly convex around the global maxima)

Local Strong Convexity and Smoothness

- We now generalize the previous setting of phase retrieval into a general local strong convexity setting:

$$\min_{\theta \in \mathbb{R}^d} f(\theta),$$

where f is locally μ-strongly convex and L-smooth in the ball $\mathbb{B}(\theta^*, r)$, i.e.,

$$\mu \leq \lambda_{\min}(\nabla^2 f(\theta)) \leq \lambda_{\max}(\nabla^2 f(\theta)) \leq L,$$

for all $\theta \in \mathbb{B}(\theta^*, r)$

- Note that, for local strong convexity, it also happens to many machine learning models, such as logistic regression

Convergence rate: Local Strong Convexity and Smoothness

Theorem 2: Assume that the function f is locally μ-strongly convex and L-smooth in $\mathbb{B}(\theta^*, r)$ where $r = \|\theta^0 - \theta^*\|$, θ^0 is an initialization, and θ^* is the minimizer of f. As long as $\eta_t = \eta \leq \frac{1}{L}$, we obtain that

$$\|\theta^t - \theta^*\| \leq \left(1 - \eta \cdot \mu\right)^{\frac{t}{2}} \|\theta^0 - \theta^*\|$$

Proof sketch: Use idea of the proof of Theorem 1, we can prove that

$$\|\theta^{t+1} - \theta^*\|^2 \leq (1 - \eta \cdot \mu)\|\theta^t - \theta^*\|^2$$

as long as $\theta^t \in \mathbb{B}(\theta^*, r)$.

- It suggests that $\theta^{t+1} \in \mathbb{B}(\theta^*, r)$ and we obtain the conclusion of the theorem

Example: Local Strong Convexity and Smoothness

- We now recall the phase retrieval problem earlier where the loss function $\overline{\mathscr{L}}_n$ is locally strongly convex
- Assume that θ_n^0 is the local initialization of $\overline{\mathscr{L}}_n$ and $\theta_n^0 \in \mathbb{B}(\widehat{\theta}_n, r_0)$ where r_0 is some given small radius and $\widehat{\theta}_n$ is an optimal solution of $\overline{\mathscr{L}}_n$
- There exist some constants μ_1 and L_1 such that $\overline{\mathscr{L}}_n$ is locally μ_1-strongly convex and L_1-smooth in $\mathbb{B}(\widehat{\theta}_n, r_0)$
- Denote by $\{\theta_n^t\}_{t\geq 0}$ the sequence of gradient descent updates for solving the loss $\overline{\mathscr{L}}_n$

Example: Local Strong Convexity and Smoothness

- From Theorem 2:

$$\|\theta_n^t - \widehat{\theta}_n\| \leq (1 - \eta \cdot \mu_1)^{t/2} \|\theta_n^0 - \widehat{\theta}_n\|$$

- Therefore, we also obtain

$$\|\theta_n^t - \theta^*\| \leq (1 - \eta \cdot \mu_1)^{t/2} \|\theta_n^0 - \widehat{\theta}_n\| + \|\widehat{\theta}_n - \theta^*\|$$

- Furthermore, we also can derive other results of phase retrieval problem as the strong convexity setting of linear regression problem

Polyak-Lojasiewicz (PL) Condition

Beyond local convergence: Polyak-Lojasiewicz (PL) Condition

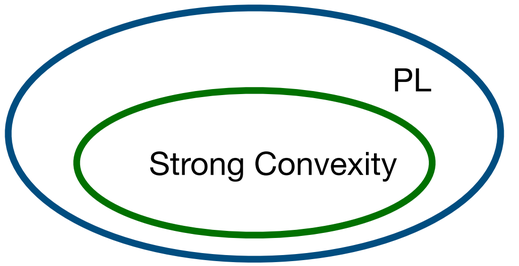
- An issue with local convergence is the local initialization
- We would like to have some guarantee beyond local convergence even when the loss function is not globally strongly convex and smooth
- An important notion is Polyak-Lojasiewicz (PL) condition

Definition 1 (PL condition): Assume that f is the loss function. The PL condition entails that there exists $\mu > 0$ such that

$$\|\nabla f(\theta)\|^2 \geq 2\mu(f(\theta) - f(\theta^*)),$$

for all $\theta \in \mathbb{R}^d$ where θ^* is the minimizer of f

Polyak-Lojasiewicz (PL) Condition



- If the function is strongly convex, it satisfies the PL condition

- The PL condition guarantees that:
    - All stationary points are global minimizers (not need to be unique)
    - It mimics the condition on the lower bound of the smallest eigenvalue of the Hessian matrix
    - This condition also motivates the Polyak average step size adaptive gradient descent (We will study it later)

Polyak-Lojasiewicz (PL) Condition

Theorem 3: Assume that the function f satisfies the PL condition for some $\mu > 0$. Furthermore, f is L-smooth. As long as $\eta_t = \eta = \frac{1}{L}$, we obtain that

$$|f(\theta^t) - f(\theta^*)| \leq \left(1 - \frac{\mu}{L}\right)^t |f(\theta^0) - f(\theta^*)|$$

- Theorem 2 guarantees global linear convergence of the objective value from any initialization θ^0

Polyak-Lojasiewicz (PL) Condition

- **Proof of Theorem 2:** Since f is L-smooth, we have

$$f(\theta^t) \leq f(\theta^{t-1}) + \langle \nabla f(\theta^{t-1}), \theta^t - \theta^{t-1} \rangle + \frac{L}{2}\|\theta^t - \theta^{t-1}\|^2$$

$$= f(\theta^{t-1}) - \eta\|\nabla f(\theta^{t-1})\|^2 + \frac{\eta^2 L}{2}\|\nabla f(\theta^{t-1})\|^2$$

$$= f(\theta^{t-1}) - \frac{1}{2L}\|\nabla f(\theta^{t-1})\|^2 \qquad \textbf{(5)}$$

Polyak-Lojasiewicz (PL) Condition

- An application of equation (5) and PL condition leads to

$$
\begin{aligned}
f(\theta^t) - f(\theta^*) &\leq f(\theta^{t-1}) - f(\theta^*) - \frac{1}{2L}\|\nabla f(\theta^{t-1})\|^2 \\
&\leq f(\theta^{t-1}) - f(\theta^*) - \frac{\mu}{L}(f(\theta^{t-1}) - f(\theta^*)) \\
&= \left(1 - \frac{\mu}{L}\right)(f(\theta^{t-1}) - f(\theta^*))
\end{aligned}
$$

- Repeating the above argument, we obtain the conclusion

Example: Polyak-Lojasiewicz (PL) Condition

- We consider the high dimensional regression problem, i.e., generalized linear model when $g(x) = x$ and $n < d$ (the number of features is more than the number of samples)

- The least-square loss is:

$$\min_{\theta \in \mathbb{R}^d} \mathscr{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - X_i^\top \theta\}^2$$

- $\nabla^2 \mathscr{L}_n(\theta) = \frac{2}{n} X^\top X \in \mathbb{R}^{d \times d}$, which is low rank since $n < d$, where

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{pmatrix} \in \mathbb{R}^{n \times d}$$

- It means that $\lambda_{\min}(\nabla^2 \mathscr{L}_n(\theta)) = 0$

Example: Polyak-Lojasiewicz (PL) Condition

- We will demonstrate that $\mathscr{L}_n$ satisfies the PL condition
- In fact, if we call $\widehat{\theta}_n$ as the optimal solution of $\mathscr{L}_n$, we have $\mathscr{L}_n(\widehat{\theta}_n) = 0$ (due to the over-parametrized condition)
- Direct calculation shows $\nabla \mathscr{L}_n(\theta) = \frac{2}{n} X^\top (X\theta - Y)$ where $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n$
- Hence, $\|\nabla \mathscr{L}_n(\theta)\|^2 = \frac{4}{n^2}(X\theta - Y)^\top XX^\top (X\theta - Y)$

$$\geq \frac{4}{n^2}\lambda_{\min}(XX^\top)\|X\theta - Y\|^2 = \frac{4\lambda_{\min}(XX^\top)}{n}(\mathscr{L}_n(\theta) - \mathscr{L}_n(\widehat{\theta}_n))$$

Example: Polyak-Lojasiewicz (PL) Condition

- It indicates that the loss function $\mathscr{L}_n$ satisfies PL condition with constant $\dfrac{4\lambda_{\min}(XX^\top)}{n}$
- Furthermore, $\mathscr{L}_n$ is $\dfrac{2\lambda_{\max}(X^\top X)}{n}$-smooth
- An application of Theorem 3 leads to

$$|\mathscr{L}_n(\theta_n^t) - \mathscr{L}_n(\widehat{\theta}_n)| \leq \left(1 - \frac{2\lambda_{\min}(XX^\top)}{\lambda_{\max}(X^\top X)}\right)^t |\mathscr{L}_n(\theta_n^0) - \mathscr{L}_n(\widehat{\theta}_n)|,$$

where $\{\theta_n^t\}_{t\geq 0}$ is a sequence of gradient descent iterates

Example: Polyak-Lojasiewicz (PL) Condition

- Since there are multiple global minimum of the least-square loss, it implies that
    - The updates from gradient descent converge to the closest minimum of the initialization

- The PL condition is **uniform**, i.e., $\|\nabla f(\theta)\|^2 \geq 2\mu(f(\theta) - f(\theta^*))$, for all $\theta \in \mathbb{R}^d$

    - This condition can be strong for several complex machine learning models, such as reinforcement learning models

    - We can adapt the PL condition to hold non-uniformly: $\|\nabla f(\theta)\|^2 \geq 2\mu(\theta)(f(\theta) - f(\theta^*))$, for all $\theta \in \mathbb{R}^d$ where $\mu(\theta)$ is some function of θ and develop normalized version of gradient descent to account for the non-uniformity (See paper [2] for detailed development)

Beyond Linear Convergence: Convex Settings

Beyond Linear Convergence

- Thus far, we have the linear convergence of the gradient descent under one of the following settings:
    - Global strong convexity and smoothness
    - Local strong convexity and smoothness
    - PL condition and smoothness
- There are ample settings that we do not have linear convergence of the gradient descent

Beyond Linear Convergence

- We consider the loss of phase retrieval problem in one dimension:

$$\min_{\theta \in \mathbb{R}} \overline{\mathscr{L}}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - (X_i \cdot \theta)^2 \}^2$$

- Recall that, $Y_i = (X_i \theta^*)^2 + \varepsilon_i$ where $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- Assume that $\theta^* = 0$
- We will demonstrate that $\overline{\mathscr{L}}_n$ is not locally strongly convex around the global minima

Beyond Linear Convergence

- Indeed, $\overline{\mathscr{L}}_n(\theta) = (\frac{1}{n}\sum_{i=1}^{n} X_i^4)\theta^4 - (\frac{2}{n}\sum_{i=1}^{n} Y_i X_i^2)\theta^2 + \frac{1}{n}\sum_{i=1}^{n} Y_i^2,$
- Different from the setting when $\theta^* \neq 0$, $\theta = 0$ has positive probability to be the global minima of $\overline{\mathscr{L}}_n$
- Indeed, since $\theta^* = 0$, we have $Y_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$

Beyond Linear Convergence

- Therefore, with high probability

$$|\frac{1}{n}\sum_{i=1}^{n} X_i^4 - 3| = \mathcal{O}(1/\sqrt{n}),$$

$$|\frac{2}{n}\sum_{i=1}^{n} Y_i X_i^2| = \mathcal{O}(1/\sqrt{n}),$$

$$|\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - 1| = \mathcal{O}(1/\sqrt{n})$$

(The proofs for these results are skipped)

Beyond Linear Convergence

- Based on these concentration inequalities, we can treat $\overline{\mathscr{L}}_n(\theta)$ as follows:

$$\overline{\mathscr{L}}_n(\theta) \approx 3\theta^4 - \frac{\theta^2}{\sqrt{n}} + 1 \quad \text{when } \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^2 < 0,$$

$$\text{and} \quad \overline{\mathscr{L}}_n(\theta) \approx 3\theta^4 + \frac{\theta^2}{\sqrt{n}} + 1 \quad \text{when } \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^2 > 0$$

- Note that, the probability $\frac{1}{n}\sum_{i=1}^{n} Y_i X_i^2 < 0$ is equal to the probability $\frac{1}{n}\sum_{i=1}^{n} Y_i X_i^2 > 0$, which is $1/2$

Beyond Linear Convergence

- Therefore, we can think that

$$\overline{\mathscr{L}}_n(\theta) \approx 3\theta^4 - \frac{\theta^2}{\sqrt{n}} + 1 \quad \text{with probability 1/2,} \quad \textbf{(C.1)}$$

$$\text{and} \quad \overline{\mathscr{L}}_n(\theta) \approx 3\theta^4 + \frac{\theta^2}{\sqrt{n}} + 1 \quad \text{with probability 1/2} \quad \textbf{(C.2)}$$

- In Case (C.2), the optimal solution is $\widehat{\theta}_n = 0$
- In Case (C.1), the optimal solutions are $\widehat{\theta}_n \approx \pm\, n^{-1/4}$ while $\theta = 0$ is local maxima

Beyond Linear Convergence

- In Case (C.1), $\nabla^2 \overline{\mathscr{L}}_n(\theta) \approx 12\theta^2 - \frac{2}{\sqrt{n}}$
- It demonstrates that the second derivative around global minima in Case (C.1) is $\mathcal{O}(1/\sqrt{n})$
- Similarly, the second derivative around global minima in Case (C.2) is also $\mathcal{O}(1/\sqrt{n})$
- These bounds on second derivatives go to 0 as $n \to \infty$
- This phenomenon suggests that the loss function is no longer locally strongly convex around the global minima as $n \to \infty$
- Hence, we may not have linear convergence of gradient updates for this case

Global Convexity

- The loss function of phase retrieval when $\theta^* = 0$ is an example of locally convex (but not locally strongly convex) loss function
- We first study the global convexity settings before discussing the local convexity settings
- We say that the function f is convex function if

$$f(\theta) \geq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle$$

for all $\theta, \theta' \in \mathbb{R}^d$

- Another definition is that $\lambda_{\min}(\nabla^2 f(\theta)) \geq 0$ for all $\theta \in \mathbb{R}^d$

Global Convexity

- When $f(\theta) = \|\theta\|^4$, we have

$$\nabla^2 f(\theta) = 8\theta\theta^\top + 4 \cdot \text{diag}(\theta_1^2, \ldots, \theta_d^2),$$

which is a semi-positive definite matrix for all $\theta \in \mathbb{R}^d$

- Therefore, f is convex function

- With similar argument, if $f(\theta) = \|\theta\|^{2p}$ where $p \geq 2$ is a positive integer, then f is a convex function

Global Convexity

Theorem 4: Assume that the function f is convex and L-smooth. As long as $\eta_t = \eta = \frac{1}{L}$, we obtain that

$$f(\theta^t) - f(\theta^*) \leq \frac{2L\|\theta^0 - \theta^*\|^2}{t}$$

Recall that, for strongly convex and smooth setting (Corollary 1), we have

$$f(\theta^t) - f(\theta^*) \leq \frac{L}{2}\left(1 - \eta \cdot \mu\right)^t \|\theta^0 - \theta^*\|^2$$

It suggests that: (i) for convex and smooth settings, the gradient descent obtains an ε-accuracy within $\mathcal{O}(1/\varepsilon)$ number of iterations

(ii) for strongly convex and smooth settings, we have $\mathcal{O}(\log(1/\varepsilon))$

Global Convexity

Proof of Theorem 4: From the convexity assumption, we have

$$f(\theta^*) - f(\theta^t) \geq \langle \nabla f(\theta^t), \theta^* - \theta^t \rangle \geq -\|\nabla f(\theta^t)\| \|\theta^t - \theta^*\|$$

Therefore,

$$\|\nabla f(\theta^t)\| \geq \frac{f(\theta^t) - f(\theta^*)}{\|\theta^t - \theta^*\|} \quad \textbf{(6)}$$

Global Convexity

Lemma 2: $\|\theta^{t+1} - \theta^*\|^2 \leq \|\theta^t - \theta^*\|^2$

- Equation (2) entails that

$$\|\theta^{t+1} - \theta^*\|^2 = \|\theta^t - \theta^*\|^2 - 2\eta\langle \nabla f(\theta^t), \theta^t - \theta^*\rangle + \eta^2\|\nabla f(\theta^t)\|^2$$

- Convexity condition means

$$f(\theta^*) \geq f(\theta^t) + \langle \nabla f(\theta^t), \theta^* - \theta^t\rangle$$

- Combining these results:

$$\|\theta^{t+1} - \theta^*\|^2 \leq \|\theta^t - \theta^*\|^2 - 2\eta(f(\theta^t) - f(\theta^*)) + \eta^2\|\nabla f(\theta^t)\|^2$$

Global Convexity

- Using the argument of Lemma 1,

$$\|\nabla f(\theta^t)\|^2 \leq 2L(f(\theta^t) - f(\theta*))$$

- Therefore, we have

$$\|\theta^{t+1} - \theta*\|^2 \leq \|\theta^t - \theta*\|^2 - \frac{\eta}{L}\|\nabla f(\theta^t)\|^2 + \eta^2\|\nabla f(\theta^t)\|^2$$

$$= \|\theta^t - \theta*\|^2$$

Global Convexity

- Applying Lemma 2 to equation (6), we have

$$\|\nabla f(\theta^t)\| \geq \frac{f(\theta^t) - f(\theta*)}{\|\theta^0 - \theta*\|}$$

- From equation (5),

$$f(\theta^{t+1}) - f(\theta^t) \leq -\frac{1}{2L}\|\nabla f(\theta^t)\|^2$$

- Combining these two inequalities:

$$f(\theta^{t+1}) - f(\theta^t) \leq -\frac{1}{2L}\frac{(f(\theta^t) - f(\theta*))^2}{\|\theta^0 - \theta*\|^2}$$

Global Convexity

- Denote $\delta_t = f(\theta^t) - f(\theta^*)$. Then, we have

$$\delta_{t+1} - \delta_t \leq -\frac{1}{2L}\frac{\delta_t^2}{\|\theta^0 - \theta^*\|^2}$$

- Dividing both sides by $\delta_t \cdot \delta_{t+1}$:

$$\frac{1}{\delta_{t+1}} \geq \frac{1}{\delta_t} + \frac{1}{2L\|\theta^0 - \theta^*\|^2}\frac{\delta_t}{\delta_{t+1}}$$

- Since $\delta_t \geq \delta_{t+1}$, we have $\frac{1}{\delta_{t+1}} \geq \frac{1}{\delta_t} + \frac{1}{2L\|\theta^0 - \theta^*\|^2}$

- Repeating this argument, we obtain the conclusion of Theorem 4

Local Convexity and Smoothness

- The general theory with globally convex function can be also adapted to locally convex function

- The function f is locally convex around the global minimum θ^* if $\lambda_{\min}(\nabla^2 f(\theta)) \geq 0$ for all $\theta \in \mathbb{B}(\theta^*, r)$ for some radius $r > 0$

Proposition 1: Assume that the function f is locally convex and L-smooth in $\mathbb{B}(\theta^*, r)$ where $r = \|\theta^0 - \theta^*\|$, θ^0 is an initialization, and θ^* is the minimizer of f. As long as $\eta_t = \eta = \frac{1}{L}$, we obtain that

$$f(\theta^t) - f(\theta^*) \leq \frac{2L\|\theta^0 - \theta^*\|^2}{t}$$

Is the rate $1/t$ tight?

- Recall our approximations for the least-square loss $\overline{\mathscr{L}}_n$ of phase retrieval in page 57:

$$\overline{\mathscr{L}}_n(\theta) \approx 3\theta^4 - \frac{\theta^2}{\sqrt{n}} + 1 \quad \text{with probability 1/2,}$$

$$\text{and} \quad \overline{\mathscr{L}}_n(\theta) \approx 3\theta^4 + \frac{\theta^2}{\sqrt{n}} + 1 \quad \text{with probability 1/2}$$

- When $n \to \infty$, we denote

$$\overline{\mathscr{L}}(\theta) := 3\theta^4 + 1$$

as the population version of the least-square loss $\overline{\mathscr{L}}_n$

Is the rate $1/t$ tight?

- The function $\overline{\mathscr{L}}$ is locally convex and 36-smooth, i.e., $L = 36$, in $\mathbb{B}(\theta^*, 1)$ where $\theta^* = 0$ is the global minimum of $\overline{\mathscr{L}}$
- Theorem 4 suggests that if $\{\theta^t\}_{t \geq 1}$ is a sequence of gradient descent updates for solving $\overline{\mathscr{L}}$, then we have

$$\overline{\mathscr{L}}(\theta^t) - \overline{\mathscr{L}}(\theta^*) \leq \frac{72 \|\theta^0 - \theta^*\|^2}{t}$$

- Unfortunately, the rate $\frac{1}{t}$ on the convergence rate of objective value of $\overline{\mathscr{L}}$ is **not tight**

Is the rate $1/t$ tight?

Claim 1: We have $\overline{\mathscr{L}}(\theta^t) - \overline{\mathscr{L}}(\theta^*) \leq \frac{C}{t^2}$ where $C > 0$ is some universal constant

Proof of Claim 1: Indeed, $\theta^t = \theta^{t-1} - \eta \nabla \overline{\mathscr{L}}(\theta^{t-1}) \ = \theta^{t-1}(1 - 12\eta(\theta^{t-1})^2)$

- We assume $\eta = \frac{1}{12}$ for the simplicity of the argument
- The above recursive equation leads to

$$\theta^t = \theta^0 \prod_{i=0}^{t-1} (1 - (\theta^i)^2)$$

Is the rate $1/t$ tight?

- It shows that $|\theta^t| \leq (1-(\theta^t)^2)^t \approx 1 - t(\theta^t)^2$
- Therefore, $|\theta^t| \approx \frac{1}{\sqrt{t}}$
- Now, $\overline{\mathscr{L}}(\theta^t) - \overline{\mathscr{L}}(\theta^*) = (\theta^t)^4 \approx \frac{1}{t^2}$ (We finish the proof of Claim 1)

Is the rate $1/t$ tight?

- The rates of objective values in Theorem 4 and Proposition 1 can be thought as **worst possible rate** when we have do not know much about the structure of function f
- In practice, we should be very careful when we use these worst rates result
    - Too pessimistic behaviors of gradient descent in the convex settings
    - Much larger sample complexity (e.g., the worst rates suggest 1 million samples but we instead only need 1000 samples for error)
    - Misleading computational efficiency (the method indeed converges faster)

Sample Complexity of Gradient Descent under Convex Settings

Sample complexity under convex settings

- We have only studied the convergence of gradient descent for solving $\overline{\mathscr{L}}$, a population version of $\overline{\mathscr{L}}_n$ (See Page 68)
- However, $\overline{\mathscr{L}}_n$ is the real objective function that we solve in practice
- We will use the approximation of $\overline{\mathscr{L}}_n$ to illustrate the difficulty of obtaining **tight** sample complexity of gradient descent method under convex settings
- To simplify the discussion, we use

$$\overline{\mathscr{L}}_n(\theta) \approx 3\theta^4 - \frac{\theta^2}{\sqrt{n}} + 1$$

with probability 1/2 (The other case can be argued in similar fashion)

Sample complexity under convex settings

- Denote θ_n^t the gradient descent updates for the approximation of $\overline{\mathscr{L}}_n$, i.e.,

$$\theta_n^t = \theta_n^{t-1} - \eta \nabla \overline{\mathscr{L}}_n(\theta) \approx \theta_n^{t-1} - \eta \left(12(\theta_n^{t-1})^3 - \frac{2\theta_n^{t-1}}{\sqrt{n}} \right)$$

- Recall that, the gradient updates θ^t for the population $\overline{\mathscr{L}}$ of $\overline{\mathscr{L}}_n$ are:

$$\theta^t = \theta^{t-1}(1 - 12\eta(\theta^{t-1})^2)$$

- The main difference between θ_n^t and θ^t is an extra noise term $\frac{\theta_n^{t-1}}{\sqrt{n}}$ (up to some fixed constant)

Sample complexity under convex settings

Theorem 5: Given the phase retrieval problem in Page 53 where $\theta^* = 0$ and the updates of gradient descent $\{\theta_n^t\}_{t\geq 0}$ in Page 75, we have

$$\|\theta_n^t - \theta^*\| \leq C \cdot n^{-1/4}$$

as long as $t \geq C' \cdot \sqrt{n}$ where C and C' are some universal constants

- The detailed proof of Theorem 5 is quite complicated; therefore, it is omitted (refer to Example 2 in Paper [3] for more detailed analysis)

Sample complexity under convex settings

A few remarks on Theorem 5:

- We need at least $\sqrt{n}$ number of iterations for this particular convex setting
    - In practice, we need to run several iterations (at least thousands) for the gradient descent updates to reach the radius $n^{-1/4}$ around the true parameter
- The radius of convergence $n^{-1/4}$ is optimal radius, i.e., the gradient descent updates obtain similar sample complexity as the global solution of phase retrieval problem

Sample complexity under convex settings

	Radius of convergence (Sample complexity)	Iteration Complexity (Computational efficiency)
Locally strongly convex and smoothness	$\|\theta_n^t - \theta^*\| \leq C \cdot n^{-1/2}$	$t \geq C' \log(n)$
Phase retrieval setting when $\theta^* = 0$	$\|\theta_n^t - \theta^*\| \leq c \cdot n^{-1/4}$	$t \geq c' \sqrt{n}$

Sample complexity under convex settings

Research Questions:

- (Q.1) For any fixed t and general convex statistical models, what is a *tight upper bound* for $\|\theta_n^t - \theta^*\|$? **(Open Question)**
- (Q.2) What is a *tight lower bound* on t such that $\|\theta_n^t - \theta^*\| \leq c \cdot \|\widehat{\theta}_n - \theta^*\|$ where $\widehat{\theta}_n$ is a global solution of the loss function? **(Open Question)**
- (Q.3) Under certain convex statistical models, can we guarantee *global convergence* of gradient descent methods? **(Open Question)**
- (Q.4) For fixed t, what is the *limiting distribution* of θ_n^t as $n \to \infty$? **(Open Question)**

Escape Saddle Points with Gradient Descent Method

$\epsilon-$Approximate Stationary Points

- Thus far, we have studied the convergence of gradient descent under (local) strongly convex or just convex settings
- The local convergence requires proper local initialization around the global minima, which can be non-trivial to design in non-convex settings
- Another useful criteria for the convergence of gradient descent is to reach a stationary point of the loss function, i.e., $\nabla f(\theta) = 0$
- Therefore, we can use the stopping criteria $\|\nabla f(\theta)\| \leq \epsilon$ ($\epsilon-$*approximate stationary point*) for the gradient descent method

$\epsilon-$Approximate Stationary Points

Theorem 6: Assume that the function f is L-smooth and $\eta_t = \eta = \frac{1}{L}$. Then, we have $\|\nabla f(\theta^t)\| \leq \epsilon$ as long as

$$t \geq \frac{2L(f(\theta^0) - f(\theta^*))}{\epsilon^2}$$

- The result demonstrates that gradient descent can find an $\varepsilon-$approximate stationary point within $\mathcal{O}(1/\varepsilon^2)$ number of iterations

$\epsilon-$Approximate Stationary Points

Proof of Theorem 6: Since f is L-smooth, from equation (5)

$$f(\theta^{t+1}) - f(\theta^t) \leq -\frac{1}{2L}\|\nabla f(\theta^t)\|^2$$

It demonstrates that

$$\frac{1}{2L}\sum_{i=0}^{t}\|\nabla f(\theta^i)\|^2 \leq \sum_{i=0}^{t}(f(\theta^i) - f(\theta^{i+1}))$$

$$= f(\theta^0) - f(\theta^{t+1}) \leq f(\theta^0) - f(\theta^*)$$

$\epsilon-$Approximate Stationary Points

Therefore, we have

$$\frac{t}{2L}\min_{0\leq i\leq t}\|\nabla f(\theta^i)\|^2 \leq f(\theta^0)-f(\theta^*)$$

$$\Rightarrow \quad \min_{0\leq i\leq t}\|\nabla f(\theta^i)\|^2 \leq \frac{2L(f(\theta^0)-f(\theta^*))}{t}$$

Since we need to find the worst possible t such that $\|\nabla f(\theta^t)\| \leq \epsilon$, we just only need to guarantee that

$\frac{2L(f(\theta^0)-f(\theta^*))}{t} \leq \epsilon^2$, which leads to $t \geq \frac{2L(f(\theta^0)-f(\theta^*))}{\epsilon^2}$

We obtain the conclusion of Theorem 6

$\epsilon-$Approximate Stationary Points

- When f is convex, finding $\varepsilon-$approximate stationary points is equivalent to finding an approximate global minimum

- We indeed have better complexity of finding $\varepsilon-$approximate stationary points when f is convex

Theorem 7: Assume that the function f is convex and L-smooth. Then, by choosing $\eta_t = \eta = \frac{1}{L}$, we have $\|\nabla f(\theta^t)\| \leq \epsilon$ as long as

$$t \geq \frac{4L\|\theta^0 - \theta^*\|}{\epsilon}$$

$\epsilon-$Approximate Stationary Points

Proof of Theorem 7: Since f is convex, using the result of Theorem 4

$$f(\theta^t) - f(\theta^*) \leq \frac{2L\|\theta^0 - \theta^*\|^2}{t}$$

Using the same proof argument from Theorem 6, we obtain

$$\frac{1}{2L}\sum_{i=t/2}^{t}\|\nabla f(\theta^i)\|^2 \leq \sum_{i=t/2}^{t}(f(\theta^i) - f(\theta^{i+1})) \leq f(\theta^{t/2}) - f(\theta^*)$$

$$\leq \frac{4L\|\theta^0 - \theta^*\|^2}{t}$$

$\epsilon-$Approximate Stationary Points

Therefore, we find that

$$\frac{t}{4L}\min_{0\le i\le t}\|\nabla f(\theta^i)\|^2 \le \frac{4L\|\theta^0-\theta^*\|^2}{t}$$

$$\Rightarrow \min_{0\le i\le t}\|\nabla f(\theta^i)\| \le \frac{4L\|\theta^0-\theta^*\|^2}{t}$$

By choosing $\frac{4L\|\theta^0-\theta^*\|^2}{t} \le \epsilon$, i.e., $t \ge \frac{4L\|\theta^0-\theta^*\|^2}{\epsilon}$, we can guarantee that $\min_{0\le i\le t}\|\nabla f(\theta^i)\| \le \epsilon$ (The conclusion of Theorem 7 follows)

Saddle Point

- In non-convex settings, getting close to $\epsilon-$approximate stationary points may be still not desirable
- $\nabla f(\theta) = 0$ still implies that θ can be a saddle point
- We say that θ is a saddle point if $\lambda_{\min}(\nabla^2 f(\theta)) \leq 0$

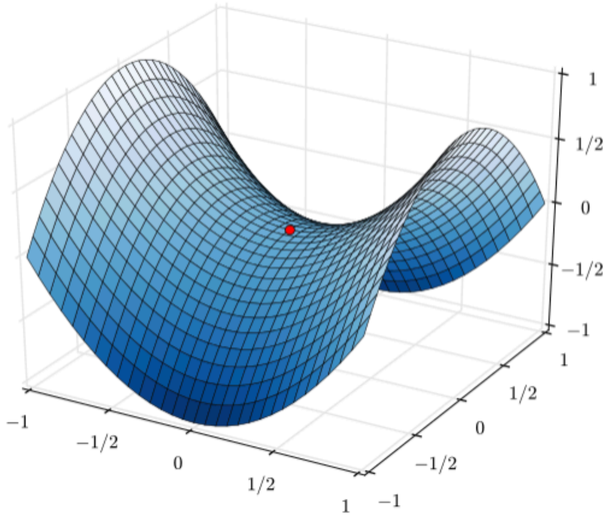


Image taken from wikipedia

For this simple illustration in 2 dimensions, the red point is local minimum in one direction, but local maxima in another direction

Escape Saddle Points

- In general, standard gradient descent may get trapped to saddle points
- There are two popular approaches for escaping saddle points with gradient descent:
    - Random initialization [5]
    - Adding noise to updates at each step [6]

Gradient Descent with Random Initialization

- Assume that we randomly initialize θ^0 from some probability distribution μ, such as Gaussian distribution

- Then, we run gradient descent: $\theta^{t+1} = \theta^t - \eta \nabla f(\theta^t)$

Theorem 8: Assume that the function f is L-smooth and $\bar{\theta}$ is a strict saddle point, i.e., $\lambda_{\min}(\nabla^2 f(\bar{\theta})) < 0$. Then, by choosing $0 < \eta < \frac{1}{L}$, we have

$$\mathbb{P}_{\mu}(\lim_{t \to \infty} \theta^t = \bar{\theta}) = 0$$

- The proof of Theorem 8 uses Manifold Stable Theorem from Dynamical System theory (Detailed proof can be found in [5])

Gradient Descent with Random Initialization

- To understand the idea of Theorem 8, we consider a simple two dimensional objective function

$$f(\theta_1, \theta_2) = \frac{1}{2}\theta_1^2 + \frac{1}{4}\theta_2^4 - \frac{1}{2}\theta_2^2$$

- Direct calculation yields:

$$\nabla f(\theta_1, \theta_2) = (\theta_1, \theta_2^3 - \theta_2)$$

- Solving $\nabla f(\theta_1, \theta_2) = 0$ yields three stationary points $\bar{\theta} = (0,0), \widehat{\theta} = (0,1), \tilde{\theta} = (0, -1)$

- We can check that $\bar{\theta}$ is strict saddle point while $\widehat{\theta}$ and $\tilde{\theta}$ are global minima

Gradient Descent with Random Initialization

- If we run gradient descent with initialization at the form $(\theta_1^0, \theta_2^0) = (z,0)$ for any $z \in \mathbb{R}$, then gradient descent updates will converge to the saddle point $\bar{\theta} = (0,0)$
- If we use different initialization of (θ_1^0, θ_2^0), the gradient descent iterates with sufficiently small step size converge to the global minima $\widehat{\theta} = (0,1)$ or $\tilde{\theta} = (0, -1)$
- Fortunately, the set $\{(z,0) : z \in \mathbb{R}\}$ has zero Lebesgue measure and with random initialization, we will avoid this set with high probability
- Therefore, with high probability gradient descent with random initialization will only converge to global minima

References

[1] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* 2020

[2] Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, Dale Schuurmans. *Leveraging Non-uniformity in First-order Non-convex Optimization.* ICML, 2021

[3] Nhat Ho, Koulik Khamaru, Raaz Dwivedi, Martin J. Wainwright, Michael I. Jordan, Bin Yu. *Instability, Computational Efficiency and Statistical Accuracy.* Arxiv Preprint, 2020

[4] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. *Gradient Descent Converges to Minimizers.* COLT, 2016

[5] Rong Ge, Furong Huang, Chi Jin, Yang Yuan. *Escaping from saddle points—online stochastic gradient for tensor decomposition.* COLT, 2015