Introduction to Machine Learning
Fall 2019
University of Science and Technology of China

Lecturer: Jie Wang | Homework 2
Posted: Oct. 05, 2019 | Due: Oct. 12, 2019
Name: San Zhang | ID: PBXXXXXXXX

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Lipschitz Continuity 10pts

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and the gradient of f is Lipschitz continuous, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \forall\, x, y \in \mathbb{R}^n,$$

where $L > 0$ is the Lipschitz constant. Please find the relation between L and the largest eigenvalue of $\nabla^2 f(x)$.

Solution: Let the largest eigenvalue of $\nabla^2 f(x)$ be $\lambda_{\max}(x)$. We show that $L > 0$ is the Lipschitz constant if and only if

$$L \geq \lambda_{\max}(x), \forall x \in \mathbb{R}^n.$$

$\Rightarrow$: For all $d \in \mathbb{R}^n$, let $x_t = x + td$, $t > 0$. Then

$$\begin{aligned} t\int_0^t \langle \nabla^2 f(x + \tau d)d, d\rangle \mathrm{d}\tau &= \langle \nabla f(x_t) - \nabla f(x), x_t - x\rangle \\ &\leq \|\nabla f(x_t) - \nabla f(x)\|_2 \|x_t - x\|_2 \\ &\leq t^2 L\|d\|_2, t > 0. \end{aligned}$$

Dividing by t^2 and letting $t \to 0^+$, we have

$$d^\top \nabla^2 f(x) d \leq L\|d\|_2^2.$$

Let d be a corresponding eigenvector of $\lambda_{\max}(x)$, and then we have

$$\lambda_{\max}(x) \leq L.$$

$\Leftarrow$: Let $L \geq \sup_x \lambda_{\max}(x)$ and suppose $L < +\infty$. For all $x, y \in \mathbb{R}^n$, let $d = y - x$. Note that $\lambda_{\max}(x) = \|\nabla^2 f(x)\|_2$. Thus we have

$$\begin{aligned} \|\nabla f(y) - \nabla f(x)\|_2 &= \| \int_0^1 \nabla^2 f(x + \tau d)d \ \mathrm{d}\tau\|_2 \\ &\leq \int_0^1 \|\nabla^2 f(x + \tau d)\|_2 \|d\|_2 \ \mathrm{d}\tau = \int_0^1 \lambda_{\max}(x + \tau d)\|d\|_2 \ \mathrm{d}\tau \\ &\leq L\|d\|_2 = L\|y - x\|_2. \end{aligned}$$

∎

Exercise 2: Gradient Descent for Convex Optimization Problems 20pts

Consider the following problem

$$\min_x f(x), \tag{1}$$

where f is convex and its gradient is Lipschitz continuous with constant $L > 0$. Assume that f can attain its minimum.

1. Show that the optimal set $\mathcal{C} = \{y : f(y) = \min_x f(x)\}$ is convex.
2. Suppose that $d(x, \mathcal{C}) = \inf_{z \in \mathcal{C}} \|x - z\|_2$. Consider the problem (1) and the sequence generated by the gradient descent algorithm. Show that $d(x_k, \mathcal{C}) \to 0$ as $k \to \infty$.

Solution:

1. Suppose $f^* = \min_x f(x)$. Let $x, y \in \mathcal{C}$. For all $\lambda \in [0, 1]$, we have

$$\begin{aligned} f^* &\leq f(\lambda x + (1-\lambda)y) \\ &\leq \lambda f(x) + (1-\lambda) f(y) \\ &= f^*. \end{aligned}$$

   Thus, $\lambda x + (1-\lambda)y \in \mathcal{C}$ and $\mathcal{C}$ is convex.

2. First, we show that $\{x_k\}$ is bounded. Let $x^* \in \mathcal{C}$. The Cosine theorem implies that

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_{k+1} - x_k + x_k - x^*\|^2 \\ &= \|x_{k+1} - x_k\|^2 + 2\langle x_{k+1} - x_k, x_k - x^* \rangle + \|x_k - x^*\|^2 \\ &= \alpha^2 \|\nabla f(x_k)\|^2 - 2\alpha \langle \nabla f(x_k), x_k - x^* \rangle + \|x_k - x^*\|^2 \\ &\leq \alpha^2 \|\nabla f(x_k)\|^2 - 2\alpha (f(x_k) - f^*) + \|x_k - x^*\|^2 \\ &\leq \frac{4}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2. \end{aligned} \tag{2}$$

   By summing up the inequality (2) for $i = 0, 1, \ldots, k$, we have

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &\leq \|x_0 - x^*\|^2 + \frac{4}{L^2} \sum_{i=0}^{k} \|\nabla f(x_i)\|^2 \\ &\leq \|x_0 - x^*\|^2 + \frac{4}{L^2} \sum_{i=0}^{\infty} \|\nabla f(x_i)\|^2 \\ &\leq \|x_0 - x^*\|^2 + \frac{4}{L^2} \frac{f(x_0) - f^*}{\alpha(1 - \frac{L}{2}\alpha)} \end{aligned}$$

   i.e., the sequence $\{x_n\}$ is bounded.

   Next, we show that $d(x_k, \mathcal{C}) \to 0$, i.e. $\limsup_k d(x_k, \mathcal{C}) = 0$.

   As $d(x_k, \mathcal{C}) \leq d(x_k, x^*)$, $d(x_k, \mathcal{C})$ is bounded and $\limsup_k d(x_k, \mathcal{C}) < +\infty$. We show that $\limsup_k d(x_k, \mathcal{C}) = 0$ by contradiction.

Suppose that $\limsup_k d(x_k, \mathcal{C}) = \epsilon > 0$. There exists a subsequence $\{x_{n_k}\}$ such that $d(x_{n_k}, \mathcal{C}) \to \epsilon$ as $k \to \infty$. Hence, there exists $K > 0$ such that $d(x_{n_k}, \mathcal{C}) \geq \frac{\epsilon}{2}$ for all $k \geq K$. As $\{x_n\}$ is bounded, $\{x_{n_k}\}$ has a further convergent subsequence $\{x_{n_{k_l}}\}$ such that $x_{n_{k_l}} \to x^\infty$ as $l \to \infty$. Clearly, $x^\infty \notin \mathcal{C}$, as $d(x_{n_k}, \mathcal{C}) \geq \frac{\epsilon}{2}$ for all $k \geq K$.

However, the continuity of f implies that

$$\begin{aligned} f^* &= \lim_{n\to\infty} f(x_n) \\ &= \lim_{l\to\infty} f(x_{n_{k_l}}) \\ &= f(x^\infty), \end{aligned}$$

i.e. $x^\infty \in \mathcal{C}$, which leads to a contradiction.

■

Exercise 3: Gradient Descent for Strongly Convex Optimization Problems 50pts

A function f is strongly convex with parameter μ if $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

1. Show that a continuously differentiable function f is strongly convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2, \forall\, x, y \in \mathbb{R}^n$$

2. Suppose that f is twice differentiable. Please find the relation between μ and the smallest eigenvalue of $\nabla^2 f(x)$.

Consider the following problem

$$\min_x f(x), \tag{3}$$

where f is strongly convex with convexity parameter $\mu > 0$ and its gradient is Lipschitz continuous with constant $L > 0$.

3. Show that the problem (3) admits a unique solution.

4. Show that

$$f(y) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2, \forall\, x, y. \tag{4}$$

5. Consider the problem (3) and the sequence generated by the gradient descent algorithm. Suppose that x^* is the solution to the problem 3. Show that

$$f(x_k) - f(x^*) \leq (1 - \mu\alpha(2 - L\alpha))^k (f(x_0) - f(x^*)).$$

   Find the range of α such that the function values $f(x_k)$ converge linearly to $f(x^*)$.

Solution:

1. Let $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$, then g is convex if and only if

$$g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle, \forall x, y \in \mathbb{R}^n.$$

   As $\nabla g(x) = \nabla f(x) - \mu x$, we get f is strongly convex if and only if

$$\begin{aligned} & g \text{ is convex,} \\ \Leftrightarrow\, & f(y) - \frac{\mu}{2}\|y\|_2^2 \geq f(x) - \frac{\mu}{2}\|x\|_2^2 + \langle \nabla f(x) - \mu x, y - x \rangle, \forall, x, y \in \mathbb{R}^n, \\ \Leftrightarrow\, & f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2, \forall x, y \in \mathbb{R}^n, \\ \Leftrightarrow\, & f \text{ is strongly convex.} \end{aligned}$$

2. Let the smallest eigenvalue of $\nabla^2 f(x)$ be $\lambda_{\min}(x)$. We show that $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex if and only if $\mu \leq \lambda_{\min}(x)$ for all $x \in \mathbb{R}^n$.

$\Rightarrow$: The strong convexity implies

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2, \\ f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|y - x\|_2^2. \end{aligned}$$

Adding them together leads to

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu \|y - x\|_2^2. \tag{5}$$

For all $d \in \mathbb{R}^n$, let $x_t = x + td$, $t > 0$. Then

$$\begin{aligned} t \int_0^t \langle \nabla^2 f(x + \tau d) d, d \rangle d\tau &= \langle \nabla f(x_t) - \nabla f(x), x_t - x \rangle \\ &\geq t^2 \mu \|d\|_2, t > 0, \end{aligned}$$

the last step comes from the inequality (5). Dividing by t^2 and letting $t \to 0^+$, we have

$$d^\top \nabla^2 f(x) d \geq \mu \|d\|_2^2.$$

Let d be a corresponding eigenvector of $\lambda_{\min}(x)$, and then we have

$$\lambda_{\min}(x) \geq \mu.$$

$\Leftarrow$: For all $z \in \mathbb{R}^n$, we have

$$\nabla^2 f(z) = U \Sigma U^\top, \quad \text{(eigen-decomposition)}$$

where $\Sigma = (\lambda_1, \lambda_2, ..., \lambda_n)$, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq \mu > 0$ and $U^\top U = I$.

For all $x, y \in \mathbb{R}^n$, there exists $t \in [0, 1]$ such that $z = tx + (1 - t)y$ and

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y - x)^\top \nabla^2 f(z)(y - x) \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y - x)^\top U \Sigma U^\top (y - x) \\ &\geq \frac{\lambda_n}{2}\|U^\top (y - x)\|_2 \\ &\geq \frac{\mu}{2}\|y - x\|_2. \end{aligned}$$

3. Existence: We have

$$\lim_{\|x\| \to +\infty} f(x) = +\infty,$$

since

$$f(x) \geq f(0) + \langle \nabla f(0), x \rangle + \frac{\mu}{2}\|x\|_2^2.$$

Thus, there exists $R > 0$ such that $f(x) \geq f(0)$ for all $x \notin B_R(0)$, where $B_R(x) = \{y : \|y - x\| \leq R\}$. Since $B_R(0)$ is compact, the continuous function $f(x)$ attains the minimum in $B_R(0)$, i.e.,

$$\exists x^* \in B_R(0), \text{ s.t. } f(x^*) = \min_{x \in B_R(0)} f(x). \quad \text{(Weierstrass's Theorem)}$$

For $x \notin B_R(0)$. we have

$$f(x) \geq f(0) \geq f(x^*)$$

Therefore, x^* is a global minimum point.

Uniqueness: Suppose that there exist $a, b \in \mathbb{R}^n$ such that $f(a) = f(b) = \min_{x \in \mathbb{R}} f(x)$ and $a \neq b$. Thus $\nabla f(a) = 0$ and

$$f(b) \geq f(a) + \frac{\mu}{2}\|b - a\|_2^2 > f(a),$$

which leads to a contradiction.

4. We have

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 \\ &= f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2 + \frac{\mu}{2}\|y - x + \frac{1}{\mu}\nabla f(x)\|_2^2 \\ &\geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2. \end{aligned}$$

5. Since the gradient of f is Lipschitz continuous, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}||y - x||_2^2.$$

The update rule implies that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \alpha(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|_2^2 \\ f(x_{k+1}) - f^* &\leq f(x_k) - f^* - \alpha(1 - \frac{L\alpha}{2})\|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - f^* - 2\mu\alpha(1 - \frac{L\alpha}{2})(f(x_k) - f^*) \\ &= (1 - \mu\alpha(2 - L\alpha))(f(x_k) - f^*), \end{aligned}$$

where the last inequality comes from the inequality (4). Therefore,

$$f(x_k) - f(x^*) \leq (1 - \mu\alpha(2 - L\alpha))^k (f(x_0) - f(x^*)).$$

To guarantee convergence, let $1 - \mu\alpha(2 - L\alpha) < 1$, i.e., $0 < \alpha < \frac{2}{L}$.

■

Exercise 4: Programming Exercise 20pts

We provide you with a data set, where the number of samples n is 16087 and the number of features d is 10013. Suppose that $x \in \mathbb{R}^{n\times d}$ is the input feature matrix and $y \in \mathbb{R}^n$ is the corresponding response vector. We use the linear model to fit the data, and thus we can formulate the optimization problem as

$$\arg\min_{\mathbf{w}} \frac{1}{n}\|y - \bar{x}\mathbf{w}\|_2^2, \tag{6}$$

where $\bar{x} = (\mathbf{1}, x) \in \mathbb{R}^{n\times(d+1)}$ and $\mathbf{w} = (w_0, w_1, \ldots, w_n)^\top \in \mathbb{R}^{d+1}$. Finish the following exercises by programming. You can use your favorite programming language.

1. Normalize the columns x_i of $\bar{x}$ $(2 \le i \le d+1)$ as follows:

   $$x_{ij} \leftarrow \frac{x_{ij} - \min(x_i)}{\max(x_i) - \min(x_i)},$$

   where x_{ij} denote thes jth entry of x_i. Use the normalized $\bar{x}$ in the following exercises.

2. Use the closed form solution to solve the problem (6), and get the solution $\mathbf{w}_0^*$.

3. Use the gradient descent algorithm to solve the problem (6). Stop the iteration until $|f(\mathbf{w}_k) - f(\mathbf{w}_0^*)| < 0.1$, where $f(\mathbf{w}) = \frac{1}{n}\|y - \bar{x}\mathbf{w}\|_2^2$. Plot $f(\mathbf{w}_k)$ versus the iteration step k.

Compare the time cost of the two approaches in 2 and 3.

Solution: Please refer to "HW2.ipynb". ■