

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

Lecturer: Jie Wang
Posted: Oct. 05, 2019
Name: Bowen Zhang

Homework 2
Due: Oct. 12, 2019
ID: PB17000215

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

Exercise 1: Lipschitz Continuity 10pts

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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:

构造函数 $g(t) = \nabla f(x + cty)$, $\forall x, y \in \mathbb{R}^n$, 其中 c 是常数。

那么存在 $\xi \in (0, 1)$, 使得:

$$\begin{aligned} \nabla f(x + cy) - \nabla f(x) &= g(1) - g(0) \\ &= g'(\xi)(1 - 0) \\ &= \nabla^2 f(x + c\xi y)cy \end{aligned}$$

两边同时取模得:

$$\Rightarrow \|\nabla^2 f(x + c\xi y)cy\|_2 = \|\nabla f(x + cy) - \nabla f(x)\|_2$$

由题目条件得:

$$\Rightarrow \|\nabla^2 f(x + c\xi y)y\|_2 \leq L\|y\|_2$$

令 $c \rightarrow 0$ 得:

$$\Rightarrow \|\nabla^2 f(x)y\|_2 \leq L\|y\|_2$$

由于对 $\nabla^2 f(x)$ 的最大特征值 λ_{max} 和对应的特征向量 y_m 有: $\nabla^2 f(x)y_m = \lambda_{max}y_m$, 综上所述对 $\nabla^2 f(x)$ 的最大特征值小于等于 L . ■

Exercise 2: Gradient Descent for Convex Optimization Problems 20pts

Consider the following problem

$$\min_x f(x), \quad (1)$$

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 $d(x_k, \mathcal{C}) \rightarrow 0$ as $k \rightarrow \infty$.

Solution:

1.

取 $y_1, y_2 \in \mathcal{C}$, 满足

$$f(y_1) = \min_x f(x);$$

$$f(y_2) = \min_x f(x).$$

考察 $\theta y_1 + (1 - \theta)y_2$, 由 $f(x)$ 是凸函数可知:

$$\begin{aligned} f(\theta y_1 + (1 - \theta)y_2) &\leq \theta f(y_1) + (1 - \theta)f(y_2) \\ &= \min_x f(x) \end{aligned}$$

由上可知 \mathcal{C} 是凸集。

2.

取 $x^* \in \mathcal{C}$, 考察:

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

其中, 用到了 $x_{k+1} = x_k - \alpha \nabla f(x_k)$.

结合 $f(x)$ 为凸函数的一阶性质:

$$f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle$$

$$\Rightarrow \text{原式} \leq 2\alpha(f(z) - f(x_k)) \leq 0.$$

课上已经求得：

$$\sum_{i=1}^{\infty} \|x_{i+1} - x_i\|^2 \leq \frac{2L}{2-L\alpha} \|f(x_0) - f^*\|^2 \dots (1)$$

现在取 x_k 的一个趋向 z 子列，其中 $z \in \mathcal{C}$ ，记为 x_{l_k} 。任取 $\delta > 0$ ，那么必定存在 l_{k_0} 使得 $\|x_{l_{k_0}} - z\|^2 \leq \frac{\delta}{2}$ ，根据 (1) 存在 l_{k_1} 使得 $\sum_{i=l_{k_1}}^{\infty} \|x_{i+1} - x_i\|^2 \leq \frac{\delta}{2}$ 。那么对于

$k > \max(k_0, k_1)$

$$\|x_k - z\|^2 \leq \|x_{l_k} - z\|^2 + \sum_{j=l_k}^{k-1} \|x_{j+1} - x_j\|^2$$

$$\leq \frac{\delta}{2} + \sum_{j=l_k}^{\infty} \|x_{j+1} - x_j\|^2$$

$$\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

令 $k \rightarrow \infty, \delta \rightarrow 0$ 即得 x_k 收敛于 z ，而 $z \in \mathcal{C}$ ，所以 $d(x_k, \mathcal{C}) \rightarrow 0$ 。

■

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

$$\min_x f(x), \quad (2)$$

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 (2) admits a unique solution.
4. Show that

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

5. Consider the problem (2) and the sequence generated by the gradient descent algorithm. Suppose that x^* is the solution to the problem 2. 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.

” \Rightarrow ”

由已知可得 $f(x) - \frac{\mu}{2}\|x\|_2^2$ 是凸函数, 那么对 $\forall x, y \in \mathbb{R}^n$, 由凸函数的一阶性质有:

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

” \Leftarrow ”

由已知, 对 $\forall x, y \in \mathbb{R}^n$ 可得:

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

而这说明 $f(x) - \frac{\mu}{2}\|x\|_2^2$ 是凸函数。

2.

取 $\forall x, y \in \mathbb{R}^n$ 有:

$$\begin{aligned}
f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 \dots (1) \\
f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2 \dots (2) \\
(1) + (2) &\Rightarrow \\
\langle \nabla f(x) - \nabla f(y), x - y \rangle &\geq \mu\|x - y\|_2^2 \dots (3)
\end{aligned}$$

构造函数 $g(t) = \nabla f(x + cty)$, 其中 c 是常数。

那么存在 $\xi \in (0, 1)$, 使得:

$$\begin{aligned}
\nabla f(x + cy) - \nabla f(x) &= g(1) - g(0) \\
&= g'(\xi)(1 - 0) \\
&= \nabla^2 f(x + c\xi y)cy \dots (4)
\end{aligned}$$

根据(3)有:

$$\langle \nabla f(x + cy) - \nabla f(x), cy \rangle \geq \mu\|cy\|_2^2 \dots (5)$$

(4)代入(5), 两边同时取模得:

$$\|\nabla^2 f(x + c\xi y)\|_2 \|y\|_2^2 \geq \mu\|y\|_2^2$$

令 $c \rightarrow 0$ 可得:

$$\|\nabla^2 f(x)\|_2 \|y\|_2^2 \geq \mu\|y\|_2^2$$

这说明对 $\nabla^2 f(x)$ 的任何特征值的绝对值都大于等于 μ , 所以其最小特征值也大于等于 μ .

3.

假设存在 x_1, x_2 满足:

$$f(x_1) = f(x_2) = \min_x f(x)$$

由 2. 中的式 (3) 可知:

$$\langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2 \rangle \geq \mu \|x_1 - x_2\|_2^2$$

由于都到达了最小值, 所以 $\nabla f(x_1) = \nabla f(x_2) = 0$, 所以有:

$$\begin{aligned} \|x_1 - x_2\|_2^2 &\leq 0 \\ \Rightarrow x_1 &= x_2 \end{aligned}$$

由此可知只有唯一解。

4.

由已知条件:

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

令 $t = \|x - y\|$, 不等式右侧看作关于 t 的二次函数, 在 $t = \frac{\|\nabla f(x)\|_2}{2 \cdot \frac{\mu}{2}}$ 时取得最小值, 于是:

$$f(y) \geq f(x) - \|\nabla f(x)\|_2 \frac{\|\nabla f(x)\|_2}{\mu} + \frac{\mu}{2} \frac{\|\nabla f(x)\|_2^2}{\mu^2}$$

于是有:

$$f(y) \geq f(x) - \frac{1}{2\mu} \|\nabla f(x)\|_2^2$$

5.

由 4. 可得:

$$\begin{aligned} f(x^*) &\geq f(x_k) - \frac{1}{2\mu} \|\nabla f(x_k)\|_2^2 \\ \Rightarrow \|\nabla f(x_k)\|_2^2 &\geq -2\mu (f(x^*) - f(x_k)) \cdots (1) \end{aligned}$$

由课上所讲的引理（从函数梯度是 Lipschitz 连续可推）有：

$$f(x_{k+1}) \leq f(x_k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x_k)\|_2^2$$

将(1)代入：

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + 2\mu\alpha \left(1 - \frac{L\alpha}{2}\right) (f(x_k) - f(x^*)) \\ &= (1 - \mu\alpha(2 - L\alpha))f(x_k) + \mu\alpha(2 - L\alpha)f(x^*) \\ f(x_{k+1}) - f(x^*) &\leq (1 - \mu\alpha(2 - L\alpha)) (f(x_k) - f(x^*)) \end{aligned}$$

不等式两边同时求和并变换下标得：

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

6.

根据线性收敛的定义，存在实数 $0 < q < 1$ ，使得 $\lim_{k \rightarrow \infty} \frac{\|f(x_{k+1}) - f(x^*)\|}{\|f(x_k) - f(x^*)\|} = q$ 则有：

$$\begin{aligned} \frac{f(x_{k+1}) - f(x^*)}{f(x_k) - f(x^*)} &\leq 1 - \mu\alpha(2 - L\alpha) < 1 \\ &\Rightarrow \alpha(2 - L\alpha) < 1 \\ &\Rightarrow 0 < \alpha < \frac{2}{L} \end{aligned}$$

■

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 $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the input feature matrix and $\mathbf{y} \in \mathbb{R}^n$ is the corresponding response vector. We use the linear model to fit the data, and thus we can formulate the optimization problem as

$$\arg \min_{\mathbf{w}} \frac{1}{n} \|\mathbf{y} - \bar{\mathbf{X}}\mathbf{w}\|_2^2, \quad (3)$$

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

1. Normalize the columns \mathbf{x}_i of $\bar{\mathbf{X}}$ ($2 \leq i \leq d+1$) as follows:

$$\mathbf{x}_{ij} \leftarrow \frac{\mathbf{x}_{ij} - \min(\mathbf{x}_i)}{\max(\mathbf{x}_i) - \min(\mathbf{x}_i)},$$

where \mathbf{x}_{ij} denote the j th entry of \mathbf{x}_i . Use the normalized $\bar{\mathbf{X}}$ in the following exercises.

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

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

Solution:

第二问和第三问的代码分别附在 *prob4.2.m* 和 *prob4.3.m* 中，数据归一化操作在每段代码开始计算之前进行。

运行环境：128 G 内存，64核 Intel(R) Xeon(R) Platinum 8153 CPU @ 2.00GHz

用时：

闭式解方法：35.072 s

梯度下降法(学习率 0.6)：1min 53.533 s

梯度下降文件中将残差保存在 *cost_history.txt* 文件中，使用 *plot.py* 进行画图，保存为 *Cost-Iterationstep.png*。 ■