SI251 - Convex Optimization, Spring 2021
Homework 4

Due on May 20, 2021, 23:59 UTC+8

Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- You are required to write down all the major steps towards making your conclusions; otherwise you may obtain limited points ($\leq$ 20%) of the problem.
- Write your homework in English; otherwise you will get no points of this homework.
- Do your homework by yourself. Any form of plagiarism will lead to 0 point of this homework. If more than one plagiarisms during the semester are identified, we will prosecute all violations to the fullest extent of the university regulations, including but not limited to failing this course, academic probation, or expulsion from the university.
- If you have any doubts regarding the grading, you need to contact the instructor or the TAs within two days since the grade is announced.

I. Gradient Methods

1. A convex function $f(\cdot)$ is said to be L-smooth if

$$f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

Show the following characterizations are equivalent to L-smooth condition.

(1) For all $\boldsymbol{x}$ and $\boldsymbol{y}$ and all $0 \le \lambda \le 1$,

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \ge \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{L}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

(10 points)

(2) $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \ge \frac{1}{L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$. (10 points)

(3) $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \le L\|\boldsymbol{x}-\boldsymbol{y}\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$. (10 points)

(4) $\left\|\nabla^2 f(\boldsymbol{x})\right\|_2 \le L, \quad \forall \boldsymbol{x}$. (if f is twice continuously differentiable) (10 points)

Solution:

Please refer to Lecture7_gradient_descent_unconstrained_note.pdf for details. Let $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \ge \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{L}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ be (0).

(0) $\Leftrightarrow$ (1) follows the similar way on the slide note of μ-strongly convex.

Then we show that the statement that $f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \le \frac{1}{2L}\|f(\boldsymbol{x})\|_2^2$ can be derived from (0) by minimizing the two side of (0). By introduce a function $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})^T \boldsymbol{x}$, where $\boldsymbol{x}^* = \boldsymbol{y}$, we get (2).

(2) $\Rightarrow$ (3) follows from the Cauchy-Schwartz inequality.

(3) $\Rightarrow$ (0):

$$\begin{aligned}
f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y}-\boldsymbol{x}) &= \int_0^1 \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle \, dt && (1)\\
&\le \int_0^1 \|\nabla f(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})) - \nabla f(\boldsymbol{x})\|_2 \|\boldsymbol{y}-\boldsymbol{x}\|_2 dt && (2)\\
&\le \int_0^1 L\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 t dt && (3)\\
&= \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2. && (4)
\end{aligned}$$

(0) $\Leftrightarrow$ (4) follows the similar way on the slide note of μ-strongly convex.

2. Suppose f is strongly convex with $mI \preceq \nabla^2 f(x) \preceq MI$. Let Δx be a descent direction at x. Show that the backtracking stopping condition holds for

$$0 < t \le -\frac{\nabla f(x)^T \Delta x}{M\|\Delta x\|_2^2}$$

Use this to give an upper bound on the number of backtracking iterations. (20 points)

Solution:

The upper bound $\nabla^2 f(x) \preceq MI$ implies

$$f(x + t\Delta x) \le f(x) + t\nabla f(x)^T \Delta x + (M/2)t^2 \Delta x^T \Delta x. \tag{5}$$

Hence $f(x+t\Delta x) \le f(x) + \alpha t \nabla f(x)^T \Delta x$ if

$$t(1-\alpha)\nabla f(x)^T \Delta x + (M/2)t^2 \Delta x^T \Delta x \le 0, \tag{6}$$

i.e., the exit condition certainly holds if $0 \le t \le t_0$ with

$$t_0 = -2(1-\alpha)\frac{\nabla f(x)^T \Delta x}{M \Delta x^T \Delta x} \ge -\frac{\nabla f(x)^T \Delta x}{M \Delta x^T \Delta x}. \tag{7}$$

Assume $t_0 \le 1$, then $\beta^k t \le t_0$ for $k \ge \log \frac{1}{t_0} / \log \frac{1}{\beta}$.

II. **Subgradient Methods**

1. On the slide of Subgradient Methods, we have the Lemma 4.1,

Lemma (4.1)**.** *Projected subgradient update rule* $\boldsymbol{x}^{t+1} = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^t - \eta_t \boldsymbol{g}^t)$ *obeys*

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - 2\eta_t\big(f(\boldsymbol{x}^t) - f^{opt}\big) + \eta_t^2\|\boldsymbol{g}^t\|_2^2,$$

where $\boldsymbol{g}^t$ *is any subgradient of* f *at* $\boldsymbol{x}^t$*.*

When f is μ-strongly convex, show that Lemma 4.1 can be improved to

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq (1-\mu\eta_t)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - 2\eta_t\big(f(\boldsymbol{x}^t) - f^{opt}\big) + \eta_t^2\|\boldsymbol{g}^t\|_2^2.$$

(20 points)

Solution:

Follow the proof of Lemma 4.1 on the slide,

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq \left\|\boldsymbol{x}^t - \boldsymbol{x}^*\right\|_2^2 - 2\eta_t \left\langle \boldsymbol{x}^t - \boldsymbol{x}^*, \boldsymbol{g}^t \right\rangle + \eta_t^2 \left\|\boldsymbol{g}^t\right\|_2^2. \tag{8}$$

Since f is μ-strongly convex, then we have

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t) \geq \left\langle \boldsymbol{x}^* - \boldsymbol{x}^t, \boldsymbol{g}^t \right\rangle + \frac{\mu}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2. \tag{9}$$

Combine these two inequalities, we complete our proof.

III. **Mirror Descent**

1. When $D_\varphi(\boldsymbol{x}, \boldsymbol{z}) = \mathrm{KL}(\boldsymbol{x}\|\boldsymbol{z}) := \sum_i x_i \log\frac{x_i}{z_i}, \mathcal{C} = \Delta := \left\{\boldsymbol{x} \in \mathbb{R}_+^n \mid \sum_i x_i = 1\right\}$, and f differentiable, show that mirror descent (MD) has the following closed-form,

$$x_i^{t+1} = \frac{x_i^t \exp\left(-\eta_t \left[\nabla f\left(\boldsymbol{x}^t\right)\right]_i\right)}{\sum_{j=1}^n x_j^t \exp\left(-\eta_t \left[\nabla f\left(\boldsymbol{x}^t\right)\right]_j\right)}, \quad 1 \leq i \leq n.$$

(20 points)

Solution:

Since f is differentiable, the update rule of MD is given by

$$\boldsymbol{x}^{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{ \left\langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \right\rangle + \frac{1}{\eta_t} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^t) \right\}, \tag{10}$$

which can be reformulate as problem $\mathcal{P}$

$$\mathcal{P} : \min_{\boldsymbol{x}\in\mathcal{C}} M(\boldsymbol{x}) = \left\langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \right\rangle + \frac{1}{\eta_t} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^t). \tag{11}$$

Then the Lagrangian function can be given as

$$\mathcal{L}(\boldsymbol{x}, \lambda) = \left\langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \right\rangle + \frac{1}{\eta_t} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^t) + \lambda(\sum_i x_i - 1). \tag{12}$$

To get $\boldsymbol{x}^{t+1}$, we have

$$\frac{\partial \mathcal{L}(\boldsymbol{x}, \lambda)}{\partial \boldsymbol{x}} = 0, \tag{13}$$

which gives

$$\nabla f(\boldsymbol{x}^t) + \frac{1}{\eta_t}\frac{\partial D_\varphi(\boldsymbol{x}, \boldsymbol{x}^t)}{\partial \boldsymbol{x}} + \lambda \mathbf{1} = 0. \tag{14}$$

Then we get

$$x_i^{t+1} = \exp\left(-\eta_t\lambda - 1\right) x_i^t \exp\left(-\eta_t \left[\nabla f\left(\boldsymbol{x}^t\right)\right]_i\right). \tag{15}$$

Combining $\sum_i x_i^{t+1} = 1$, we have

$$\exp\left(-\eta_t\lambda - 1\right) = \frac{1}{\sum_{j=1}^n x_j^t \exp\left(-\eta_t \left[\nabla f\left(\boldsymbol{x}^t\right)\right]_j\right)}. \tag{16}$$

Hence,

$$x_i^{t+1} = \frac{x_i^t \exp\left(-\eta_t \left[\nabla f\left(\boldsymbol{x}^t\right)\right]_i\right)}{\sum_{j=1}^n x_j^t \exp\left(-\eta_t \left[\nabla f\left(\boldsymbol{x}^t\right)\right]_j\right)}, \quad 1 \leq i \leq n. \tag{17}$$