HW 1(a)

Group 3

February 16, 2019

Strong Convexity

0.1 Part 1

We need to prove the equivalence of the following three statements

$$\nabla^2 f(x) \succeq \mu I \tag{1}$$

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

and

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \mu \|x - y\|_2^2 \tag{3}$$

We will prove the equivalence as follows $1 \implies 2 \implies 3$

Firstly, we show that $f(x)$ is strongly convex if and only if $\nabla^2 f(x) \succ \mu I$ for any $x \in \mathcal{X}$ where f is assumed to be twice continuously differentiable.

Suppose that $\nabla^2 f(x) \geq \mu I$. We then prove the gradient inequality to prove the strong convexity of $f(x)$. To this end, we define

$$g(x) := f(x) - \frac{\mu}{2}\|x\|_2^2 \ , \tag{4}$$

where $\nabla g(x) = \nabla f(x) - \mu x$ and $\nabla^2 g(x) = \nabla^2 f(x) - \mu I$. As we assumed that $\nabla^2 f(x) \geq \mu I$, then $\nabla^2 g(x) \geq 0$.

By the linear approximation of Taylor's theorem, we have

$$g(y) = g(x) + \nabla g(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 g(z)(y - x) \ , \tag{5}$$

where $z \in [x, y]$ and $z \in \mathcal{X}$.

Since $\nabla^2 g(z) \geq 0$, it follows that $(y-x)^T \nabla^2 g(z)(y-x) \geq 0$. So, $g(y) \geq g(x) + \nabla g(x)^T (y-x)$, and hence $f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|_2^2$.

Now, we show that if $g(x)$ is convex, then $g(x)$ is positive semidefinite, and thereby $f(x)$ is positive definite.

For a small perturbed vector to x, i.e., $(x+\alpha y) \in \mathcal{X}$ for $0 < \alpha < \epsilon$, where ϵ is a small enough positive number. If we invoke the gradient inequality, then we have

$$g(x+\alpha y) \geq g(x) + \alpha \nabla g(x)^T y \ . \tag{6}$$

By the second order approximation based on Taylor's theorem, we have

$$g(x+\alpha y) = g(x) + \alpha \nabla g(x)^T y + \frac{\alpha^2}{2} y^T \nabla^2 g(x) y + o(\alpha^2 \|y\|^2) \ . \tag{7}$$

If we combine (14) with (13), then we have

$$\frac{\alpha^2}{2} y^T \nabla^2 g(x) y + o(\alpha^2 \|y\|^2) \geq 0. \tag{8}$$

Dividing the above inequality (15) by α^2 we have

$$\frac{1}{2} y^T \nabla^2 g(x) y + \frac{o(\alpha^2 \|y\|^2)}{\alpha^2} \geq 0. \tag{9}$$

To this end, $\alpha \to 0^+$, we conclude that

$$y^T \nabla^2 g(x) y \geq 0, \tag{10}$$

for any $y \in R^n$, implying that $\nabla^2 g(x) \succeq 0$.

Finally, plugging in the Hessian of $g(x)$, i.e., $\nabla^2 g(x) = \nabla^2 f(x) - \mu I$ in the above positive semidifinitess condition, we conclude that $\nabla^2 f(x) \geq \mu I$.

1 $\implies$ **2** Using taylor expansion, for any twice differential function for some x, y we have

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$$

Using $\nabla^2 f(x) \succeq \mu I$, we have

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

2 $\implies$ **3** From 2, we have

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

and

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

Adding the above two we get

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

Proving 1 from 2 and 3 concludes this exercise. This is done below

Firstly, we show that $f(x)$ is strongly convex if and only if $\nabla^2 f(x) \succ \mu I$ for any $x \in \mathcal{X}$ where f is assumed to be twice continuously differentiable. Suppose that $\nabla^2 f(x) \geq \mu I$. We then prove the gradient inequality to prove the strong convexity of $f(x)$. To this end, we define

$$g(x) := f(x) - \frac{\mu}{2}\|x\|_2^2 \ , \tag{11}$$

where $\nabla g(x) = \nabla f(x) - \mu x$ and $\nabla^2 g(x) = \nabla^2 f(x) - \mu I$. As we assumed that $\nabla^2 f(x) \geq \mu I$, then $\nabla^2 g(x) \geq 0$.

By the linear approximation of Taylor's theorem, we have

$$g(y) = g(x) + \nabla g(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 g(z)(y-x) \ , \tag{12}$$

where $z \in [x, y]$ and $z \in \mathcal{X}$.

Since $\nabla^2 g(z) \geq 0$, it follows that $(y-x)^T\nabla^2 g(z)(y-x) \geq 0$. So, $g(y) \geq g(x) + \nabla g(x)^T(y-x)$, and hence $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2$.

Now, we show that if $g(x)$ is convex, then $g(x)$ is positive semidefinite, and thereby $f(x)$ is positive definite.

For a small perturbed vector to x, i.e., $(x+\alpha y) \in \mathcal{X}$ for $0 < \alpha < \epsilon$, where ϵ is a small enough positive number. If we invoke the gradient inequality, then we have

$$g(x+\alpha y) \geq g(x) + \alpha \nabla g(x)^T y \ . \tag{13}$$

By the second order approximation based on Taylor's theorem, we have

$$g(x+\alpha y) = g(x) + \alpha \nabla g(x)^T y + \frac{\alpha^2}{2} y^T \nabla^2 g(x) y + o(\alpha^2 \|y\|^2) \ . \tag{14}$$

If we combine (14) with (13), then we have

$$\frac{\alpha^2}{2} y^T \nabla^2 g(x) y + o(\alpha^2 \|y\|^2) \geq 0. \tag{15}$$

Dividing the above inequality (15) by α^2 we have

$$\frac{1}{2} y^T \nabla^2 g(x) y + \frac{o(\alpha^2 \|y\|^2)}{\alpha^2} \geq 0. \tag{16}$$

To this end, $\alpha \to 0^+$, we conclude that

$$y^T \nabla^2 g(x) y \geq 0, \tag{17}$$

for any $y \in R^n$, implying that $\nabla^2 g(x) \succeq 0$.

Finally, plugging in the Hessian of $g(x)$, i.e., $\nabla^2 g(x) = \nabla^2 f(x) - \mu I$ in the above positive semidefiniteness condition, we conclude that $\nabla^2 f(x) \geq \mu I$.

Ques (a)

Note that for a fixed x, the above becomes a convex quadratic function of y. If we want to find the minimum over all y we differentiate. On differentiating we find that $f'(y) = 0$ when we put $\overline{y} = x - \frac{1}{\mu} \nabla f(x)$. Therefore again starting from 2, we have

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

Since the above is true for every y thus we have $f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2$

Ques (b)

In the lecture it is given that

$$(\nabla f(x_1) - \nabla f(x_2))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2$$

1(a)-(b) is immediately from the above by using cauchy schwarz inquality.

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

Hence we get

$$\|\nabla f(x_1) - \nabla f(x_2)\|_2 \geq \mu \|x_2 - x_1\|_2 \tag{18}$$

Ques (c)

For 1(a)-(c) we multiply 18 by $\|\nabla f(x_1) - \nabla f(x_2)\|_2$ on both sides and use cauchy schwarz

$$\begin{aligned} \|\nabla f(x_1) - \nabla f(x_2)\|_2^2 &\geq \mu \|x_2 - x_1\|_2 \|\nabla f(x_1) - \nabla f(x_2)\|_2 \\ &\geq (x_2 - x_1)^T (\nabla f(x_1) - \nabla f(x_2)) \end{aligned}$$

Ques (d)

let $g(x)$ be the given μ strongly convex function and h be a convex function, then we have

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

Also,

$$h(y) \geq h(x) + \nabla h(x)^T (y - x)$$

Let $f(x) = h(x) + g(x)$, then we have

$$\begin{aligned} f(y) =& g(y) + h(y) \\ &\geq g(x) + h(x) + (\nabla g(x) + \nabla h(x))^T (y - x) + \frac{\mu}{2} \|y - x\|^2 \\ &= f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2 \end{aligned}$$

Hence, f which was a sum of a strongly convex function and a convex function is indeed strongly convex. This proves **1(a)-(d)**

EP3260: Homework #1b

Submit on January 28, 2019 by Xin

Group 3

1

Given $\forall x_1, x_2 \in R^d, \| \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \|_2 \leq L \| x_2 - x_1 \|_2$, prove $f(x_2) \leq f(x_1) + \bigtriangledown f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{L}{2} \| x_2 - x_1 \|_2^2, \forall x_1, x_2$.

proof:

Follow Cauchy-Schwarz inequality: $\langle x, v \rangle \leq \| x \|_2 \| v \|_2$, we have $\langle \bigtriangledown f(x_2) - \bigtriangledown f(x_1), x_2 - x_1 \rangle \leq \| \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \|_2 \| x_2 - x_1 \|_2$. By transitivity:

$$\begin{aligned} &\langle \bigtriangledown f(x_2) - \bigtriangledown f(x_1), x - y \rangle \leq L \| x_2 - x_1 \|_2^2 \\ \Longrightarrow &(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T (x - y) \leq L \| x_2 - x_1 \|_2^2 \end{aligned} \tag{1}$$

We can find a function $g(x) = \frac{L}{2} x^T x - f(x)$. $\bigtriangledown g(x) = Lx - \bigtriangledown f(x)$.

$$\begin{aligned} &(\bigtriangledown g(x_2) - \bigtriangledown g(x_1))^T (x_2 - x_1) \\ =&(Lx_2 - \bigtriangledown f(x_2) - Lx_1 + \bigtriangledown f(x_1))^T (x_2 - x_1) \\ =&(L(x_2 - x_1) - (\bigtriangledown f(x_2) - \bigtriangledown f(x_1)))^T (x_2 - x_1) \\ \overset{(1)}{\Longrightarrow}&(2) \geq 0 \end{aligned} \tag{2}$$

Due to the monotone gradient condition for convexity[1], we know $g(x)$ is convex. Again, using the first order condition[2] of g(x), we have:

$$\begin{aligned} &g(x_2) \geq g(x_1) + \bigtriangledown g(x_1)^T (x_2 - x_1) \\ \Longrightarrow &\frac{L}{2} x_2^T x_2 - f(x_2) \geq \frac{L}{2} x_1^T x_1 - f(x_1) + (Lx_1 - \bigtriangledown f(x_1))^T (x_2 - x_1) \\ \Longrightarrow &f(x_2) \leq f(x_1) + \frac{L}{2} (x_2^T x_2 - x_1^T x_1) - (Lx_1 - \bigtriangledown f(x_1))^T (x_2 - x_1) \\ \Longrightarrow &f(x_2) \leq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{L}{2} (x_2^T x_2 - x_1^T x_1 - 2x_1^T x_2 + 2x_1^T x_1) \\ \Longrightarrow &f(x_2) \leq f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{L}{2} \| x_2 - x_1 \|_2^2 \end{aligned} \tag{3}$$

Thus, the proof is completed.

2

Given $\forall x_1, x_2 \in R^d, \| \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \|_2 \leq L \| x_2 - x_1 \|_2$, prove $f(x_2) \geq f(x_1) + \bigtriangledown f(x_1) + \bigtriangledown f(x_1)^T (x_2 - x_1) + \frac{1}{2L} \| \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \|_2^2, \forall x_1, x_2$.

proof: The proof is based on the assumption that f is convex. We need to form a function: $k_{x_1}(a) = f(a) - \bigtriangledown f(x_1)^T a$, and $k_{x_1}(a)$ get minimizer at $a^* = x_1$.

Then we can form another function: $i_{x_1}(a) = \frac{L}{2} a^T a - k_{x_1}(a)$. Since the previous $g(x)$ is convex, $i_{x_1}(a)$ is also convex, thus

$$k_{x_1}(a) \leq k_{x_1}(x_2) + \bigtriangledown k_{x_1}(x_2)^T (a - x_2) + \frac{L}{2} \| a - x_2 \|_2^2 \tag{4}$$

[1] $g(x)$ is convex if and only if $(\bigtriangledown g(x) - \bigtriangledown g(y))^T (x - y) \geq 0, \forall x, y$.

[2] $g(x)$ is convex if and only if $g(y) \geq g(x) + \bigtriangledown g(x)^T (y - x), \forall x, y$.

Problem 2 continued on next page...

$$
\begin{aligned}
& k_{x_1}(x_2) + \bigtriangledown k_{x_1}(x_2)^T(a - x_2) + \frac{L}{2} \parallel a - x_2 \parallel_2^2 \\
=& \frac{L}{2}(\parallel a - x_2 \parallel_2^2 + 2 \times \frac{1}{L} \bigtriangledown k_{x_1}(x_2)^T(a - x_2) + \frac{1}{L^2} \parallel \bigtriangledown k_{x_1}(x_2) \parallel_2^2) - \frac{1}{2L} \bigtriangledown k_{x_1}(x_2) + k_{x_1}(x_2) \\
=& \frac{L}{2} \parallel a - x_2 + \frac{1}{L} k_{x_1}(x_2) \parallel_2^2 - \frac{1}{2L} \bigtriangledown k_{x_1}(x_2) + k_{x_1}(x_2) \\
\geq& - \frac{1}{2L} \bigtriangledown k_{x_1}(x_2) + k_{x_1}(x_2) \qquad (5)
\end{aligned}
$$

Take $a = x_1$ in equation (4) we have:

$$
\begin{aligned}
f(x_2) - f(x_1) - \bigtriangledown f(x_1)^T(x_2 - x_1) &= k_{x_1}(x_2) - k_{x_1}(x_1) \\
&\overset{(5)}{\Longrightarrow} \geq \frac{1}{2L} \parallel \bigtriangledown k_{x_1}(x_2) \parallel_2^2 \\
&= \frac{1}{2L} \parallel \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \parallel_2^2 \qquad (6)
\end{aligned}
$$

Then we have $f(x_2) \geq f(x_1) + \bigtriangledown f(x_1) + \bigtriangledown f(x_1)^T(x_2 - x_1) + \frac{1}{2L} \parallel \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \parallel_2^2, \forall x_1, x_2$.

3

Given $\forall x_1, x_2 \in R^d, \parallel \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \parallel_2 \leq L \parallel x_2 - x_1 \parallel_2$, prove $(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T(x_2 - x_1) \geq \frac{1}{L} \parallel \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \parallel_2^2, \forall x_1, x_2$.

proof:

The proof is based on the assumption that f is convex.

Using the similar method in proving (b), we define $k_{x_2}(a) = f(a) - \bigtriangledown f(x_2)^T a$, and $k_{x_2}(a)$ get minimizer at $a^* = x_2$. Together with $i_{x_2}(a) = \frac{L}{2} a^T a - k_{x_2}(a)$. Since the previous $g(x)$ is convex, $i_{x_2}(a)$ is also convex, thus $k_{x_2}(a) \leq k_{x_2}(x_1) + \bigtriangledown k_{x_2}(x_1)^T(a - x_1) + \frac{L}{2} \parallel a - x_1 \parallel_2^2$.

Take $a = x_2$ then we have:

$$
\begin{aligned}
f(x_1) - f(x_2) - \bigtriangledown f(x_2)^T(x_1 - x_2) &= k_{x_2}(x_1) - k_{x_2}(x_2) \\
&\geq \frac{1}{2L} \parallel \bigtriangledown k_{x_2}(x_1) \parallel_2^2 \\
&= \frac{1}{2L} \parallel \bigtriangledown f(x_1) - \bigtriangledown f(x_2) \parallel_2^2 \qquad (7)
\end{aligned}
$$

Adding equation (6) and (7) together then we have:

$(\bigtriangledown f(x_2) - \bigtriangledown f(x_1))^T(x_2 - x_1) \geq \frac{1}{L} \parallel \bigtriangledown f(x_2) - \bigtriangledown f(x_1) \parallel_2^2, \forall x_1, x_2$.

And the proof is complete.

EP3260:Fundamentals of Machine Learning over Networks

Group 3

Notation

Upper-case letters with a double underline denotes matrices, e.g., $\boldsymbol{A}$. Lower-case letters with a single underline denotes vectors, e.g., $\boldsymbol{a}$. The i-th element of the vector $\boldsymbol{a}$ is denoted by either $a\left[i\right]$ or a_i, and element in the i-th row and j-th column of the matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}\left[i,j\right]$. An i-th column vector of a matrix $\boldsymbol{A}$ is denoted as either $\boldsymbol{A}_i$ or $\boldsymbol{A}[:,i]$.

HW1(c)- part1

Resource allocation problem is given as:

$$\begin{aligned} \underset{\{x_i\in\mathbb{R}\}}{\text{minimize}} \quad & \frac{1}{N}\sum_{i=1}^{N} f_i\left(x_i\right) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}\,, \end{aligned} \tag{1}$$

where $\boldsymbol{A}\in\mathbb{R}^{p\times N}$, $\boldsymbol{b}\in\mathbb{R}^{p\times 1}$, and $\boldsymbol{x}=[x_1,\ldots,x_N]^T$.

(a) Assume strong-convexity and smoothness on f. How would you solve this problem when $N=1000$?

Solution:

Depending on the (time/complexity) requirement and exact function, we can solve the problem via centralized or distributed first/second order optimization methods. We enlist at least three possible methods to solve the problem (1).

To this end, the problem (1) can be re-written as

$$\begin{aligned} \underset{\{x_i\in\mathbb{R}\}}{\text{minimize}} \quad & \sum_{i=1}^{N} f_i\left(x_i\right) \\ \text{subject to} \quad & \sum_{i=1}^{N} \boldsymbol{A}_i x_i=\boldsymbol{b} \quad \forall i=1,\ldots,N\,, \end{aligned} \tag{2}$$

where $\boldsymbol{A}_i\in\mathbb{R}^{p\times 1}$ corresponds to the column vector of $\boldsymbol{A}$ matrix.

1. (Optimal method: Closed-form depending on the function f_i)
   Forming Lagrangian of the problem in (2) yields

$$L\left(\{x_i\},\{\boldsymbol{\lambda}_i\}\right)=\sum_{i=1}^{N} f_i\left(x_i\right)+\sum_{i=1}^{N}\boldsymbol{\lambda}_i^T\left(\boldsymbol{A}_i x_i-\boldsymbol{b}\right), \tag{3}$$

where $\boldsymbol{\lambda}_i \in \mathbb{R}^{p\times 1}$ is a Lagrange multiplier.
Now, we can attempt to solve $\nabla_{\{x_i\},\{\boldsymbol{\lambda}_i\}} L\left(\{x_i\}, \{\boldsymbol{\lambda}_i\}\right) = 0$ and desirably obtain the optimal $\{x_i\}$ without Lagrange multipliers.

If no closed-form exists, then we naturally resort to iterative schemes. It seems that if $N = 1000$ one could employ either first or preferably second order methods depending on their requirements.

2. **Newton's Method**–Second Order Method:
Let us define

$$\overline{f}\left(\boldsymbol{x}\right) \coloneqq \sum_{i=1}^{N} f_i\left(x_i\right) \ , \tag{4}$$

such that we can equivalently repose the problem in (1) as following:

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \quad \overline{f}\left(\boldsymbol{x}\right) \\ &\text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \ . \end{aligned} \tag{5}$$

The Newton's method with an equality constraint (affine) start with $\boldsymbol{x}^{(0)} \in \mathbb{R}^N$ and then can have the following iterative scheme

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + \alpha \boldsymbol{v}, \tag{6}$$

where α is a step length and

$$\boldsymbol{v} = \underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min} \quad \overline{f}\left(\boldsymbol{x}\right) + \nabla\overline{f}\left(\boldsymbol{x}\right)^T \boldsymbol{z} + \frac{1}{2}\boldsymbol{z}^T\nabla^2\,\overline{f}\left(\boldsymbol{x}\right)\boldsymbol{z} \quad \text{subject to} \quad \boldsymbol{A}\left(\boldsymbol{x}+\boldsymbol{z}\right) = \boldsymbol{b} \ . \tag{7}$$

Now, we form the Lagrangian to the problem (7), that is

$$L\left(\boldsymbol{z}, \boldsymbol{\mu}\right) = \overline{f}\left(\boldsymbol{x}\right) + \nabla\overline{f}\left(\boldsymbol{x}\right)^T \boldsymbol{z} + \frac{1}{2}\boldsymbol{z}^T\nabla^2\,\overline{f}\left(\boldsymbol{x}\right)\boldsymbol{z} + \boldsymbol{\mu}^T\left(\boldsymbol{A}\left(\boldsymbol{x}+\boldsymbol{z}\right) - \boldsymbol{b}\right). \tag{8}$$

Based on KKT conditions, we can form the linear system combining stationarity and primal feasibility conditions, i.e.,

$$\begin{bmatrix} \nabla^2\,\overline{f}\left(\boldsymbol{x}\right) & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} -\nabla\overline{f}\left(\boldsymbol{x}\right) \\ -\left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\right) \end{bmatrix} \ . \tag{9}$$

Let us define the gradient and Hessian of $\overline{f}\left(\boldsymbol{x}\right) = \sum_{i=1}^{N} f_i\left(x_i\right)$ as

$$\boldsymbol{g} \coloneqq \nabla\overline{f}\left(\boldsymbol{x}\right) = \left[\nabla f_1\left(x_1\right), \ldots, \nabla f_N\left(x_N\right)\right]^T \tag{10}$$

$$\boldsymbol{H} \coloneqq \nabla^2\,\overline{f}\left(\boldsymbol{x}\right) = \text{Diag}\left\{\nabla^2 f_1\left(x_1\right), \ldots, \nabla^2 f_N\left(x_N\right)\right\} \ , \tag{11}$$

respectively.
Since Hessian is positive definite, i.e., $\boldsymbol{H}$ is invertible, then solve for $\boldsymbol{v}$ and $\boldsymbol{\mu}$ according to the following equations:

$$\boldsymbol{A}\boldsymbol{H}^{-1}\boldsymbol{A}^T\boldsymbol{\mu} = \left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\right) - \boldsymbol{A}\boldsymbol{H}^{-1}\boldsymbol{g} \tag{12}$$

$$\boldsymbol{H}\boldsymbol{v} = -\boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\mu} \ . \tag{13}$$

3. (Iterative) **First Order Methods**:

   First order methods can be employed which is known to offer low computational complexity per iteration since Hessian inverse is not computed, but it may take a large number of iterations to converge (depending on the function/problem). There are several first order based algorithms. The problem in hand has similarities to consensus like optimization problem, where the problem is divided into smaller problems, e.g., dual decomposition, consensus ADMM, etc.

   In this homework, we present a so-called Dual Decomposition (cf. Boyd) scheme.

   Lagrangian (3) is separable in $\boldsymbol{x}$, i.e.,

$$L\left(\{x_i\},\{\boldsymbol{\lambda}_i\}\right) = \sum_{i=1}^{N} \underbrace{\left[f_i\left(x_i\right) + \boldsymbol{\lambda}_i^T \boldsymbol{A}_i x_i\right]}_{:=L_i(x_i,\boldsymbol{\lambda}_i)} - \boldsymbol{\lambda}_i^T \boldsymbol{b} \tag{14}$$

$$= \sum_{i=1}^{N} L_i\left(x_i,\boldsymbol{\lambda}_i\right) - \boldsymbol{\lambda}_i^T \boldsymbol{b}\,. \tag{15}$$

   Dual decomposition method exploits the separable Lagrangian, which are easier to solve analytically. The resulting iterative scheme (for a kth iteration update) can be summarized as following:

$$x_i^{(k)} = \arg\min_{x_i} L_i\left(x_i,\boldsymbol{\lambda}_i^{(k-1)}\right) = \arg\min_{x_i}\left\{f_i\left(x_i\right) + \boldsymbol{\lambda}_i^T \boldsymbol{A}_i x_i\right\} = \partial f_i\left(x_i\right) + \boldsymbol{A}_i^T \boldsymbol{\lambda}_i \tag{16}$$

$$\boldsymbol{\lambda}_i^{(k)} = \boldsymbol{\lambda}_i^{(k)} + \alpha^{(k)}\left[\sum_{i=1}^{N} \boldsymbol{A}_i x_i^{(k)} - \boldsymbol{b}\right]\,. \tag{17}$$

   However, f_i are strongly convex, then the Dual decomposition method may converge but can be slow.

(b) What if $N = 10^9$?

- One could use either first order or second order depending on the requirement. Since Hessian (11) is a diagonal matrix, then the Hessian inverse is feasible for such a large scale system. So, Newton's like method can be employed.

(c) Can we use Newtons method for $N = 10^9$? Try efficient method for computing $\nabla^2 f(x_k)$ for $p = 1$ and $b = 1$ (probability simplex constraint). Extend it to $1 \leq p \ll N$.

- Yes, we can use Newton's method since Hessian inverse for Newton's method would be computationally feasible for such large-scale problems as Hessian (11) is a diagonal matrix.

  The probability simplex problem for $p = 1$ and $b = 1$ can be shown as

$$\begin{aligned} \underset{\{x_i\in\mathbb{R}\}}{\text{minimize}} \quad & \frac{1}{N}\sum_{i=1}^{N} f_i\left(x_i\right) \\ \text{subject to} \quad & \sum_{i=1}^{N} x_i = \mathbf{1}^T \boldsymbol{x} = 1 \quad \forall i = 1,\ldots,N\,, \\ & x_i \geq 0\,. \end{aligned} \tag{18}$$

If we ignore the inequality constraint on x_i, i.e.,

$$\begin{aligned} &\underset{\{x_i \in \mathbb{R}\}}{\text{minimize}} \quad \frac{1}{N} \sum_{i=1}^{N} f_i\left(x_i\right) \\ &\text{subject to} \quad \sum_{i=1}^{N} x_i = \mathbf{1}^T \boldsymbol{x} = 1 \quad \forall i = 1, \ldots, N \ , \end{aligned} \tag{19}$$

then it's an equality constrained problem and we can use Newton's method as described above from (4) to (13). Moreover, Newton's method can be employed for $1 \leq p \ll N$ since only Hessian inverse is required.

HW1(c)- part2

Now, add twice differentiable $r(x)$ to the objective in (1) such that the problem reads

$$\begin{aligned} \underset{\{x_i \in \mathbb{R}\}}{\text{minimize}} \quad & \frac{1}{N} \sum_{i=1}^{N} f_i\left(x_i\right) + r(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\,. \end{aligned} \tag{20}$$

(a) Assume strong-convexity and smoothness on f. How would you solve this problem when $N = 1000$?

Solution:

To this end, the problem (20) can be re-written as

$$\begin{aligned} \underset{\{x_i \in \mathbb{R}\}}{\text{minimize}} \quad & \sum_{i=1}^{N} f_i\left(x_i\right) + r(\boldsymbol{x}) \\ \text{subject to} \quad & \sum_{i=1}^{N} \boldsymbol{A}_i x_i = \boldsymbol{b} \quad \forall i = 1, \ldots, N\,, \end{aligned} \tag{21}$$

where $\boldsymbol{A}_i \in \mathbb{R}^{p \times 1}$ corresponds to the column vector of $\boldsymbol{A}$ matrix.

1. (Optimal method: Closed-form depending on the function f_i and r)
   Forming Lagrangian of the problem in (21) yields

$$L\left(\{x_i\}, \{\boldsymbol{\lambda}_i\}\right) = \sum_{i=1}^{N} f_i\left(x_i\right) + r(\boldsymbol{x}) + \sum_{i=1}^{N} \boldsymbol{\lambda}_i^T \left(\boldsymbol{A}_i x_i - \boldsymbol{b}\right), \tag{22}$$

   where $\boldsymbol{\lambda}_i \in \mathbb{R}^{p \times 1}$ is a Lagrange multiplier.
   If the closed-form does not exist, then similar to the part 1, we need to resort to iterative schemes.
2. **Newton's Method**–Second Order Method:
   Let us define

$$\overline{f}\left(\boldsymbol{x}\right) \coloneqq \sum_{i=1}^{N} f_i\left(x_i\right) + r(\boldsymbol{x})\,, \tag{23}$$

   such that we can equivalently repose the problem in (1) as following:

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \quad & \overline{f}\left(\boldsymbol{x}\right) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\,. \end{aligned} \tag{24}$$

   The Newton's method with an equality constraint (affine) start with $\boldsymbol{x}^{(0)} \in \mathbb{R}^N$ and then can have the following iterative scheme

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + \alpha \boldsymbol{v}, \tag{25}$$

where α is a step length and

$$\boldsymbol{v} = \underset{\boldsymbol{x}\in\mathbb{R}^N}{\arg\min} \quad \overline{f}\left(\boldsymbol{x}\right) + \nabla\overline{f}\left(\boldsymbol{x}\right)^T \boldsymbol{z} + \frac{1}{2}\boldsymbol{z}^T\nabla^2\,\overline{f}\left(\boldsymbol{x}\right)\boldsymbol{z} \quad \text{subject to} \quad \boldsymbol{A}\left(\boldsymbol{x}+\boldsymbol{z}\right) = \boldsymbol{b}\,. \tag{26}$$

Now, we form the Lagrangian to the problem (26), that is

$$L\left(\boldsymbol{z},\boldsymbol{\mu}\right) = \overline{f}\left(\boldsymbol{x}\right) + \nabla\overline{f}\left(\boldsymbol{x}\right)^T \boldsymbol{z} + \frac{1}{2}\boldsymbol{z}^T\nabla^2\,\overline{f}\left(\boldsymbol{x}\right)\boldsymbol{z} + \boldsymbol{\mu}^T\left(\boldsymbol{A}\left(\boldsymbol{x}+\boldsymbol{z}\right)-\boldsymbol{b}\right). \tag{27}$$

Based on KKT conditions, we can form the linear system combining stationarity and primal feasibility conditions, i.e.,

$$\begin{bmatrix} \nabla^2\,\overline{f}\left(\boldsymbol{x}\right) & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix}\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} -\nabla\overline{f}\left(\boldsymbol{x}\right) \\ -\left(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\right) \end{bmatrix}\,. \tag{28}$$

Let us define the gradient and Hessian of $\overline{f}\left(\boldsymbol{x}\right) = \sum_{i=1}^N f_i\left(x_i\right)$ as

$$\boldsymbol{g} \coloneqq \nabla\overline{f}\left(\boldsymbol{x}\right) = \left[\nabla f_1\left(x_1\right),\ldots,\nabla f_N\left(x_N\right)\right]^T \tag{29}$$

$$\boldsymbol{H} \coloneqq \nabla^2\,\overline{f}\left(\boldsymbol{x}\right) \overset{?}{\neq} \text{Diag}\left\{\nabla^2 f_1\left(x_1\right),\ldots,\nabla^2 f_N\left(x_N\right)\right\}, \tag{30}$$

respectively.

Since Hessian is positive definite, i.e., $\boldsymbol{H}$ is invertible, then solve for $\boldsymbol{v}$ and $\boldsymbol{\mu}$ according to the following equations:

$$\boldsymbol{A}\boldsymbol{H}^{-1}\boldsymbol{A}^T\boldsymbol{\mu} = \left(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\right) - \boldsymbol{A}\boldsymbol{H}^{-1}\boldsymbol{g} \tag{31}$$

$$\boldsymbol{H}\boldsymbol{v} = -\boldsymbol{g} - \boldsymbol{A}^T\boldsymbol{\mu}\,. \tag{32}$$

Notice that the Hessian may not be a diagonal matrix and employing Newton's method for large scale systems may computationally be infeasible.

3. (Iterative) First Order Methods:
   After including $r(\boldsymbol{x})$ in the objective, the Lagrangian is not separable so, for instance, dual decomposition cannot be employed. However, some other first order methods could be employed.

(b) What if $N = 10^9$?

- Since Hessian may not be a diagonal matrix, then the Hessian inverse would be computationally infeasible for such a large-scale problem. Thus, one needs to resort to first order methods.

(c) Can we use Newtons method for $N = 10^9$? Try efficient method for computing $\nabla^2 f(x_k)$ for $p = 1$ and $b = 1$ (probability simplex constraint). Extend it to $1 \leq p \ll N$.

- No, Newton's method is not appealing and may be computationally insfeasible as the Hessian matrix is not a diagonal matrix.
  The probability simplex problem for $p = 1$ and $b = 1$ can be shown as

$$\begin{aligned} \underset{\{x_i\in\mathbb{R}\}}{\text{minimize}} \quad & \frac{1}{N}\sum_{i=1}^N f_i\left(x_i\right) + r(\boldsymbol{x}) \\ \text{subject to} \quad & \sum_{i=1}^N x_i = \boldsymbol{1}^T\boldsymbol{x} = 1 \quad \forall i = 1,\ldots,N\,, \\ & x_i \geq 0\,. \end{aligned} \tag{33}$$

If we ignore the inequality constraint on x_i, i.e.,

$$\begin{aligned} \underset{\{x_i \in \mathbb{R}\}}{\text{minimize}} \quad & \frac{1}{N} \sum_{i=1}^{N} f_i\left(x_i\right) + r(\boldsymbol{x}) \\ \text{subject to} \quad & \sum_{i=1}^{N} x_i = \mathbf{1}^T \boldsymbol{x} = 1 \quad \forall i = 1, \ldots, N \ , \end{aligned} \tag{34}$$

then it is an equality constrained optimization problem. However, Hessian inverse for Newton's method is still a problem even for $p = 1$.

Machine Learning Over Networks
Homework Assignment 1(d)

Problem 1(d)

Assuming that $f(\mathbf{x})$ is μ-strongly convex and L-smooth, use the coercivity of the gradient to prove:

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{\mu + L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 .$$

Proof. Let us define $h(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2} \|\mathbf{x}\|_2^2$. Since $f(\mathbf{x})$ is strongly convex, this implies that $h(\mathbf{x})$ is convex. In addition, $f(\mathbf{x})$ is L-smooth, which implies that $h(\mathbf{x})$ is $(L - \mu)$-smooth.

Then, we can use the coercivity of the gradient in $h(\mathbf{x})$ as:

$$(\nabla h(\mathbf{x}) - \nabla h(\mathbf{y}))^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) \geq \frac{1}{L - \mu} \|\nabla h(\mathbf{x}) - \nabla h(\mathbf{y})\|_2^2 , \text{ using def. of } h(\mathbf{x})$$

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) - \mu (\mathbf{x} - \mathbf{y}))^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) \geq \frac{1}{L - \mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) - \mu (\mathbf{x} - \mathbf{y})\|_2^2 , \text{ using def. of 2-norm}$$

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) - \mu \|\mathbf{x} - \mathbf{y}\|_2 \geq \frac{1}{L - \mu} \left[\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 - 2\mu (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) + \mu \|\mathbf{x} - \mathbf{y}\|_2^2 \right] ,$$

$$\left(1 + \frac{2\mu}{L - \mu}\right) (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) \geq \frac{1}{L - \mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 + \left(\mu + \frac{\mu^2}{L - \mu}\right) \|\mathbf{x} - \mathbf{y}\|_2^2 , \text{rearranging terms}$$

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^{\mathrm{T}} (\mathbf{x} - \mathbf{y}) \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{L + \mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 .$$

□