EP3260: Fundamentals of Machine Learning over Networks
Homework Assignment 1, Due Date: February 19, 2020
Group 4: H. Hellström, M. Tavana, F. Vannella, and M. Zeng

Problem 1. A differentiable function f is μ-strongly convex iff $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \mu > 0$

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2. \tag{1}$$

Then proof that:

- (1) is equivalent to a minimum positive curvature $\nabla^2 f(\boldsymbol{x}) \succeq \mu \boldsymbol{I}_d, \quad \forall \boldsymbol{x} \in \mathcal{X}$;

  Proof. According to (1), we have

  $$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_1\|_2^2 + \frac{\mu}{2}\|\boldsymbol{x}_2\|_2^2 - \mu \boldsymbol{x}_1^T \boldsymbol{x}_2,$$

  or

  $$f(\boldsymbol{x}_2) - \frac{\mu}{2}\|\boldsymbol{x}_2\|_2^2 \geq f(\boldsymbol{x}_1) - \frac{\mu}{2}\|\boldsymbol{x}_1\|_2^2 + (\nabla f(\boldsymbol{x}_1) - \mu \boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1).$$

  Lets define $g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$. We have $\nabla g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \mu \boldsymbol{x}$ and $\nabla^2 g(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) - \mu \boldsymbol{I}_d$, therefore

  $$g(\boldsymbol{x}_2) \geq g(\boldsymbol{x}_1) + \nabla g(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1), \tag{2}$$

  which is the first-order condition for convexity of $g(\boldsymbol{x})$. Hence, $\nabla^2 g(\boldsymbol{x}) \succeq \boldsymbol{0}$, or equivalently, $\nabla^2 f(\boldsymbol{x}) - \mu \boldsymbol{I}_d \succeq \boldsymbol{0}$. This implies $\nabla^2 f(\boldsymbol{x}) \succeq \mu \boldsymbol{I}_d$. □

- (1) is equivalent to $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \mu \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$;

  Proof. We know that $g(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ is a convex function. It follows from the monotone gradient condition for convexity that $g(\boldsymbol{x})$ is convex if and only if $(\nabla g(\boldsymbol{x}_2) - \nabla g(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq 0$. Therefore

  $$(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1) - \mu(\boldsymbol{x}_2 - \boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq 0,$$

  or

  $$(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \mu \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2.$$

  □

- (1) implies

  (a) $f(\boldsymbol{x}) - f^* \leq \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2, \quad \forall \boldsymbol{x}$;

  Proof. By taking minimization with respect to $\boldsymbol{x}_2$ from both sides of (1), we have

  $$f^* \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2^* - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_2^* - \boldsymbol{x}_1\|_2^2, \tag{3}$$

  where $\nabla f(\boldsymbol{x}_1) + \mu(\boldsymbol{x}_2^* - \boldsymbol{x}_1) = 0$, or $\boldsymbol{x}_2^* = -\frac{1}{\mu}\nabla f(\boldsymbol{x}_1) + \boldsymbol{x}_1$. By substituting $\boldsymbol{x}_2^*$ in (3), we have

  $$\begin{aligned} f^* &\geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T \left(-\frac{1}{\mu}\nabla f(\boldsymbol{x}_1) + \boldsymbol{x}_1 - \boldsymbol{x}_1\right) + \frac{\mu}{2}\left\|-\frac{1}{\mu}\nabla f(\boldsymbol{x}_1) + \boldsymbol{x}_1 - \boldsymbol{x}_1\right\|_2^2 \\ &= f(\boldsymbol{x}_1) - \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}_1)\|_2^2, \end{aligned}$$

or

$$f(\boldsymbol{x}) - f^* \leq \frac{1}{2\mu} \|\nabla f(\boldsymbol{x})\|_2^2, \quad \forall \boldsymbol{x}.$$

□

(b) $\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \leq \dfrac{1}{\mu} \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2, \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2;$

Proof. Using Cauchy-Schwartz inequality on the equivalent condition $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \mu \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$, we have

$$\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \geq (\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \mu \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2. \tag{4}$$

Dividing both sides by $\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2$ (assuming $\boldsymbol{x}_2 \neq \boldsymbol{x}_1$) gives

$$\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 \geq \mu \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2. \tag{5}$$

□

(c) $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \leq \dfrac{1}{\mu} \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2, \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2;$

Proof. From (4) and (5) we have

$$\frac{(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1)}{\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2} \leq \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \leq \frac{1}{\mu} \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2.$$

Hence, $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \leq \dfrac{1}{\mu} \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2$. □

(d) $f(\boldsymbol{x}) + r(\boldsymbol{x})$ is strongly convex for any convex f and strongly convex r.

Proof. Assuming differentiability of $f(\boldsymbol{x})$ and $r(\boldsymbol{x})$, we have by definition of convexity of f and strong convexity of r

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1), \tag{6}$$

$$r(\boldsymbol{x}_2) \geq r(\boldsymbol{x}_1) + \nabla r(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2} \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2. \tag{7}$$

By summing (6) and (7), we have

$$f(\boldsymbol{x}_2) + r(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + r(\boldsymbol{x}_1) + \nabla(r(\boldsymbol{x}_1) + f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2} \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2,$$

which is the definition of strong convexity for $f(\boldsymbol{x}) + r(\boldsymbol{x})$. □

Problem 2. A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth iff it is differentiable and its gradient is L-Lipschitz continuous (usually w.r.t. norm-2):

$$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d, \left\|\nabla f\left(\boldsymbol{x}_2\right) - \nabla f\left(\boldsymbol{x}_1\right)\right\|_2 \leq L \left\|\boldsymbol{x}_2 - \boldsymbol{x}_1\right\|_2. \tag{8}$$

For all $\boldsymbol{x}_1, \boldsymbol{x}_2$, prove that (8) implies

(a) $f\left(\boldsymbol{x}_2\right) \leq f\left(\boldsymbol{x}_1\right) + \nabla f\left(\boldsymbol{x}_1\right)^T\left(\boldsymbol{x}_2 - \boldsymbol{x}_1\right) + \frac{L}{2}\left\|\boldsymbol{x}_2 - \boldsymbol{x}_1\right\|_2^2$;

Proof. To begin, let's define the helper function $g\left(\boldsymbol{x}\right) = \frac{L}{2}\boldsymbol{x}^T\boldsymbol{x} - f\left(\boldsymbol{x}\right)$ and prove that it is convex. $g\left(\boldsymbol{x}\right)$ is convex iff the first-order condition holds [1]

$$g\left(\boldsymbol{x}_2\right) \geq g\left(\boldsymbol{x}_1\right) + \nabla g\left(\boldsymbol{x}_1\right)^T\left(\boldsymbol{x}_2 - \boldsymbol{x}_1\right). \tag{9}$$

Of course, the variables in (9) can be switched to create

$$g\left(\boldsymbol{x}_1\right) \geq g\left(\boldsymbol{x}_2\right) + \nabla g\left(\boldsymbol{x}_2\right)^T\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right). \tag{10}$$

Taking $g\left(\boldsymbol{x}_1\right)$ from (10) and inserting it into (9) gives

$$g\left(\boldsymbol{x}_2\right) \geq \left(g\left(\boldsymbol{x}_2\right) + \nabla g\left(\boldsymbol{x}_2\right)^T\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right)\right) + \nabla g\left(\boldsymbol{x}_1\right)^T\left(\boldsymbol{x}_2 - \boldsymbol{x}_1\right),$$

and after rearranging that becomes

$$\left(\nabla g\left(\boldsymbol{x}_2\right) - \nabla g\left(\boldsymbol{x}_1\right)\right)^T\left(\boldsymbol{x}_2 - \boldsymbol{x}_1\right) \geq 0. \tag{11}$$

Inserting the definition of the helper function into (11) gives

$$\left(L\boldsymbol{x}_2 - \nabla f\left(\boldsymbol{x}_2\right) - \left(L\boldsymbol{x}_1 - \nabla f\left(\boldsymbol{x}_1\right)\right)^T\left(\boldsymbol{x}_2 - \boldsymbol{x}_1\right) \geq 0.$$

This can be rearranged to resemble (8):

$$L\left\|\boldsymbol{x}_2 - \boldsymbol{x}_1\right\|_2^2 \geq \left(\nabla f\left(\boldsymbol{x}_2\right) - \nabla f\left(\boldsymbol{x}_1\right)\right)^T\left(\boldsymbol{x}_2 - \boldsymbol{x}_1\right). \tag{12}$$

Multiplying both sides of (8) with $\left\|\boldsymbol{x}_2 - \boldsymbol{x}_1\right\|_2$ and then using Cauchy-Schwarz inequality on the L.H.S. gives (12). Thus, (8) implies (12), which in turn proves that the first-order conditions for $g\left(\boldsymbol{x}\right)$ holds. Using the now established convexity of $g\left(\boldsymbol{x}\right) = \frac{L}{2}\boldsymbol{x}^T\boldsymbol{x} - f\left(\boldsymbol{x}\right)$, we can insert it into the first-order condition (9) to show the final result: $\frac{L}{2}\boldsymbol{x}_2^T\boldsymbol{x}_2 - f\left(\boldsymbol{x}_2\right) \geq \frac{L}{2}\boldsymbol{x}_1^T\boldsymbol{x}_1 - f\left(\boldsymbol{x}_1\right) + \left(L\boldsymbol{x}_1 - \nabla f\left(\boldsymbol{x}_1\right)\right)^T\left(\boldsymbol{x}_2 - \boldsymbol{x}_1\right)$. Rearranging terms and noting that $\frac{L}{2}\left(\boldsymbol{x}_2^T\boldsymbol{x}_2 + \boldsymbol{x}_1^T\boldsymbol{x}_1\right) - L\boldsymbol{x}_1^T\boldsymbol{x}_2 = \frac{L}{2}\left\|\boldsymbol{x}_2 - \boldsymbol{x}_1\right\|_2^2$ yields $f\left(\boldsymbol{x}_2\right) \leq f\left(\boldsymbol{x}_1\right) + \nabla f\left(\boldsymbol{x}_1\right)^T\left(\boldsymbol{x}_2 - \boldsymbol{x}_1\right) + \frac{L}{2}\left\|\boldsymbol{x}_2 - \boldsymbol{x}_1\right\|_2^2$. □

(b) $f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2$;

Proof. Let us consider f is convex. Therefore, the function $\Phi_{\boldsymbol{x}_1}(\boldsymbol{z}) \triangleq f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}_1)^T \boldsymbol{z}$, will also be convex and its optimum occurs at $\boldsymbol{z}^* = \boldsymbol{x}_1$. Moreover, we know $\nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{z}) = \nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}_1)$. Hence,

$$\nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{z}) - \nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2) = \nabla f(\boldsymbol{z}) - \nabla f(\boldsymbol{x}_2). \tag{13}$$

By substituting (13) in (8), we have

$$\|\nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{z}) - \nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2)\|_2 \leq L\|\boldsymbol{z} - \boldsymbol{x}_2\|_2. \tag{14}$$

According to the part (a), (14) is equivalent to

$$\Phi_{\boldsymbol{x}_1}(\boldsymbol{z}) \leq \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2) + \nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2)^T (\boldsymbol{z} - \boldsymbol{x}_2) + \frac{L}{2}\|\boldsymbol{z} - \boldsymbol{x}_2\|_2^2.$$

Taking minimization with respect to z on both sides, yields,

$$\begin{aligned}
f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_1)^T \boldsymbol{x}_1 &\leq f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)^T \boldsymbol{x}_2 + \nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2)^T (\boldsymbol{z}^* - \boldsymbol{x}_2) + \frac{L}{2}\|\boldsymbol{z}^* - \boldsymbol{x}_2\|_2^2 \\
&= f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)^T \boldsymbol{x}_2 + \nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2)^T \left(-\frac{1}{L}\nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2)\right) + \frac{L}{2}\left\|-\frac{1}{L}\nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2)\right\|_2^2 \\
&= f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)^T \boldsymbol{x}_2 - \frac{1}{2L}\|\nabla \Phi_{\boldsymbol{x}_1}(\boldsymbol{x}_2)\|_2^2 \\
&= f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)^T \boldsymbol{x}_2 - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2
\end{aligned}$$

Re-arranging gives

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2. \tag{15}$$

□

(c) $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \frac{1}{L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2$.

Proof. Using the result from b), we know that (8) implies (15). Using (15) as a starting point, we begin by rearranging the terms to get

$$\nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \leq f(\boldsymbol{x}_2) - f(\boldsymbol{x}_1) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2. \tag{16}$$

Then, we switch $\boldsymbol{x}_2$ with $\boldsymbol{x}_1$ in (16) to get:

$$\nabla f(\boldsymbol{x}_2)^T (\boldsymbol{x}_1 - \boldsymbol{x}_2) \leq f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2. \tag{17}$$

The sum of (16) and (17) is

$$(\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \leq -\frac{1}{L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2.$$

Multiplying both sides by -1 gives the final result

$$(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \frac{1}{L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2. \tag{18}$$

To summarize, b) shows that (8) implies (15), and c) shows that (15) implies (18), thus (8) implies (18). □

Assume convexity if needed.

Problem 3. Define, discuss the benefits, and give examples for the different convergence rates of a sequence of updates $\{\boldsymbol{x_k}\}$:

First, we define a limit which will be useful to express the convergence rate of a sequence of updates $\{\boldsymbol{x_k}\}$. Let's assume that the sequence will eventually converge, so that

$$\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^*$$

holds. Then, the following expression can be used to represent how quickly the sequence converges:

$$\lim_{k\to\infty} \frac{|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*|}{|\boldsymbol{x}_k - \boldsymbol{x}^*|^p} = m. \tag{19}$$

The values of p and m determines the convergence rate of the sequence.

(a) Sublinear

Sublinear is the slowest form of convergence, and is defined by a series which converge according to (19) with $m = 1$ and $p = 1$. For simplicity, the example will be a series with scalar $\{\boldsymbol{x}_k\}$. The following sequence has sublinear convergence rate:

$$a_k = \frac{1}{k}. \tag{20}$$

This sequence converges to zero, we can verify the sublinear convergence by inserting it into (19) with $p = 1$:

$$\lim_{k\to\infty} \frac{1/(k+1) - 0}{1/k - 0} = \lim_{k\to\infty} \frac{k}{k+1} = 1.$$

(b) Linear

Linear convergence is faster than sublinear, and is defined by a series which converge according to (19) with $0 < m < 1$ and $p = 1$. Gradient descent over an L-smooth and μ-strongly convex function has a linear convergence rate with

$$m = 1 - \frac{2}{1 + L/\mu}.$$

Since $\mu < L$ by definition, we see that $0 < m < 1$ which makes the convergence linear.

(c) Superlinear

Superlinear convergence is faster than linear, and is defined by a series which converge according to (19) with $m = 0$ and $p = 1$. To find an example of a sequence with superlinear convergence, we need a series where the ratio of two adjacent updates goes to zero as k approaches infinity. An example of such a series is:

$$x_k = 1 + \left(\frac{1}{k}\right)^k.$$

This sequence converges to one, we can verify the superlinear convergence by inserting it into (19) with $p = 1$:

$$\lim_{k\to\infty} \frac{1 + 1/\left(k+1\right)^{k+1} - 1}{1 + 1/k^k - 1} = \lim_{k\to\infty} \frac{k^k}{\left(k+1\right)^{k+1}} = 0.$$

(d) Quadratic

Quadratic convergence the fastest of the four forms of convergence covered here, and is defined by a series which converge according to (19) with $m > 0$ and $p = 2$. Given that $\boldsymbol{x}_k$ is sufficiently close to the minimum, the Newton method experiences quadratic convergence. However, in practice, another algorithm has to be ran for a couple of steps before the Newton method can be applied to ensure that the current $\boldsymbol{x}_k$ is sufficiently close to the minimum.

Problem 4. Consider

$$\begin{aligned} &\text{minimize} && \frac{1}{N}\sum_{i\in[N]} f_i\left(x_i\right) \\ &\text{subject to} && A\boldsymbol{x} = \boldsymbol{b}, \end{aligned}$$

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

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

Proof. This is a convex optimization problem with equality constraint. There are a few ways to handle it:

(a) Eliminating equality constraints. This can be done by finding a matrix F and vector $\hat{\boldsymbol{x}}$ that parameterize the feasible set:

$$\{\boldsymbol{x}|A\boldsymbol{x} = \boldsymbol{b}\} = \{F\boldsymbol{z} + \hat{\boldsymbol{x}}|\boldsymbol{z} \in \mathbb{R}^{n-p}\} \tag{22}$$

Then, the problem can be transformed into

$$\min \hat{f}\left(\boldsymbol{z}\right) = f\left(F\boldsymbol{z} + \hat{\boldsymbol{x}}\right). \tag{23}$$

For the above problem, we can use gradient descent to solve it. Since f is strong-convex and smooth, we could achieve linear convergence.

(b) Solving equality constrained problems via the dual. The dual problem is given by

$$g\left(\boldsymbol{v}\right) = -\boldsymbol{b}^T\boldsymbol{v} - f^*\left(-A^T\boldsymbol{v}\right) \tag{24}$$

(c) We could solve the problem directly using Newton's method with equality constraints. In this case, the Newton step Δx_{nt} is characterized by

$$\begin{bmatrix} \nabla^2 f\left(\boldsymbol{x}\right) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{nt} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f\left(\boldsymbol{x}\right) \\ 0 \end{bmatrix} \tag{25}$$

(d) We could also solve it using ADMM. This is because the variables in the objective function is separate. Therefore, we could update one variable at one time, while fixing the others.

□

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

Proof. In this case, Newton's method cannot be applied since the matrix is too large. However, other methods, e.g., ADMM can still be applied. □

(c) Can we use Newton's method for $N = 10^9$? Try efficient method for computing $\nabla^2 f\left(\boldsymbol{x}_k\right)$ for $p = 1$ and $b = 1$ (probability simplex constraint). Extend it to $1 \leq p \ll N$.

Proof. Since x_i is separate in the objective function, the Hessian matrix $\nabla^2 f\left(\boldsymbol{x}\right)$ is a diagonal matrix, i.e.,

$$\nabla^2 f\left(\boldsymbol{x}\right) = \frac{1}{N}\begin{bmatrix} \frac{\partial^2 f_1}{\partial x_1^2} & \cdots & 0 \\ 0 & \frac{\partial^2 f_i}{\partial x_i^2} & 0 \\ 0 & \cdots & \frac{\partial^2 f_N}{\partial x_N^2} \end{bmatrix} \tag{26}$$

By exploiting the diagonal property, we may still use Newton's method. □

(d) Now, add twice differentiable $r(\boldsymbol{x})$ to the objective and solve (a)–(c).

Proof. Here it is unclear whether $r(\boldsymbol{x})$ is convex or not! If $r(\boldsymbol{x})$ is convex, the similar procedure can be performed after adding $r(\boldsymbol{x})$ for (a). However, for (b)-(c), Newton's method cannot be used since the Hessian matrix is no longer diagonal. □

Problem 5. In the convergence proof of GD with constant step size and strongly convex objective function (see slides), prove the coercivity of the gradient:

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

Proof. Let us consider the function $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$. As already noted in Problem 1 Equation (2), since $f(\boldsymbol{x})$ is $\mu-$strongly convex, the function $g(\boldsymbol{x})$ is convex, i.e.

$$g(\boldsymbol{x}_2) \geq g(\boldsymbol{x}_1) + \nabla g(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1), \tag{27}$$

Also, we observe that since $f(\boldsymbol{x})$ is L-smooth, we have that $g(\boldsymbol{x})$ is also smooth with smoothness parameter $L - \mu$:

$$\begin{aligned} \|\nabla g(\boldsymbol{x}_2) - \nabla g(\boldsymbol{x}_1)\|_2 &= \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1) - \mu(\boldsymbol{x}_2 - \boldsymbol{x}_1)\|_2 \\ &\leq \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 - \mu\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \\ &\leq (L - \mu)\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \end{aligned} \tag{28}$$

$$(\nabla g(\boldsymbol{x}_2) - \nabla g(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \frac{1}{L - \mu}\|\nabla g(\boldsymbol{x}_2) - \nabla g(\boldsymbol{x}_1)\|_2^2 \tag{29}$$

By expressing it in function of $f(\boldsymbol{x})$, we have

$$\begin{aligned} &(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1) - \mu(\boldsymbol{x}_2 - \boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \frac{1}{L - \mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1) - \mu(\boldsymbol{x}_2 - \boldsymbol{x}_1))\|_2^2 \\ &\left(1 + \frac{2\mu}{L - \mu}\right)(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \frac{1}{L - \mu}\|\nabla f(x) - \nabla f(y)\|^2 + \left(\frac{\mu^2}{L - \mu} + \mu\right)\|x - y\|^2 \\ &\left(\frac{L + \mu}{L - \mu}\right)(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \frac{1}{L - \mu}\|\nabla f(x) - \nabla f(y)\|^2 + \left(\frac{\mu L}{L - \mu}\right)\|x - y\|^2 \\ &(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T (\boldsymbol{x} - \boldsymbol{y}) \geq \frac{\mu L}{\mu + L}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{\mu + L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \end{aligned} \tag{30}$$

$\square$

References

[1] Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.