

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

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

Then proof that:

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

Proof. According to (1), we have

What about converse?

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

or

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

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

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

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

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

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

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

or

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

Converse is missing

\square

- (1) implies

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

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

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

Correct

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

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

or

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

□

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

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

Correct

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

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

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

□

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

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

Correct

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

$$\text{Hence, } (\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1))^T (\mathbf{x}_2 - \mathbf{x}_1) \leq \frac{1}{\mu} \|\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1)\|_2^2.$$

□

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

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

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

Correct

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

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

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

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

□

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

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

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

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

Correct

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

$$g(\mathbf{x}_2) \geq g(\mathbf{x}_1) + \nabla g(\mathbf{x}_1)^T (\mathbf{x}_2 - \mathbf{x}_1). \quad (9)$$

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

$$g(\mathbf{x}_1) \geq g(\mathbf{x}_2) + \nabla g(\mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2). \quad (10)$$

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

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

and after rearranging that becomes

$$(\nabla g(\mathbf{x}_2) - \nabla g(\mathbf{x}_1))^T (\mathbf{x}_2 - \mathbf{x}_1) \geq 0. \quad (11)$$

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

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

This can be rearranged to resemble (8):

$$L \|\mathbf{x}_2 - \mathbf{x}_1\|_2^2 \geq (\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1))^T (\mathbf{x}_2 - \mathbf{x}_1). \quad (12)$$

Multiplying both sides of (8) with $\|\mathbf{x}_2 - \mathbf{x}_1\|_2$ 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 $g(\mathbf{x})$ holds.

Using the now established convexity of $g(\mathbf{x}) = \frac{L}{2} \mathbf{x}^T \mathbf{x} - f(\mathbf{x})$, we can insert it into the first-order

condition (9) to show the final result: $\frac{L}{2} \mathbf{x}_2^T \mathbf{x}_2 - f(\mathbf{x}_2) \geq \frac{L}{2} \mathbf{x}_1^T \mathbf{x}_1 - f(\mathbf{x}_1) + (L\mathbf{x}_1 - \nabla f(\mathbf{x}_1))^T (\mathbf{x}_2 - \mathbf{x}_1)$.

Rearranging terms and noting that $\frac{L}{2} (\mathbf{x}_2^T \mathbf{x}_2 + \mathbf{x}_1^T \mathbf{x}_1) - L\mathbf{x}_1^T \mathbf{x}_2 = \frac{L}{2} \|\mathbf{x}_2 - \mathbf{x}_1\|_2^2$ yields $f(\mathbf{x}_2) \leq f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^T (\mathbf{x}_2 - \mathbf{x}_1) + \frac{L}{2} \|\mathbf{x}_2 - \mathbf{x}_1\|_2^2$. \square

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

Correct

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

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

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

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

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

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

Taking minimization with respect to \mathbf{z} on both sides, yields,

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

Re-arranging gives

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

□

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

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

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

Correct

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

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

The sum of (16) and (17) is

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

Multiplying both sides by -1 gives the final result

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

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

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

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

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}^*$$

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

$$\lim_{k \rightarrow \infty} \frac{|\mathbf{x}_{k+1} - \mathbf{x}^*|}{|\mathbf{x}_k - \mathbf{x}^*|^p} = m. \quad (19)$$

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 $\{x_k\}$. The following sequence has sublinear convergence rate:

$$a_k = \frac{1}{k}. \quad (20)$$

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

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

(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 $p = 1$:

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

(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 \mathbf{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 \mathbf{x}_k is sufficiently close to the minimum.

Correct

Problem 4. Consider

Correct

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

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

- (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 $\hat{\mathbf{x}}$ that parameterize the feasible set:

$$\{\mathbf{x} | A\mathbf{x} = \mathbf{b}\} = \{F\mathbf{z} + \hat{\mathbf{x}} | \mathbf{z} \in \mathbb{R}^{n-p}\} \quad (22)$$

Then, the problem can be transformed into

$$\min \hat{f}(\mathbf{z}) = f(F\mathbf{z} + \hat{\mathbf{x}}). \quad (23)$$

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

$$g(\mathbf{v}) = -\mathbf{b}^T \mathbf{v} - f^*(-A^T \mathbf{v}) \quad (24)$$

- (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

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

- (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 $\nabla^2 f(\mathbf{x}_k)$ 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 $\nabla^2 f(\mathbf{x})$ is a diagonal matrix, i.e.,

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

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



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

Proof. Here it is unclear whether $r(\mathbf{x})$ is convex or not! If $r(\mathbf{x})$ is convex, the similar procedure can be performed after adding $r(\mathbf{x})$ 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:

Correct

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

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

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

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

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

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

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

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

□



References

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