Homework 1

Ali Bemani **Ali.Bemani@hig.se**
Oscar Bautista Gonzalez **Oscar.Bautista.Gonzalez@hig.se**
Martin Hellkvist **Martin.Hellkvist@angstrom.uu.se**
Syed Aqeel Raza **s.aqeelraza@gmail.com**

Solutions are provided in formatted form for problems 1.1, 1.3 and 1.4. Due to time limitations we decided to not format the solutions for 1.2 and 1.5. These solutions are provided within this document, as photocopies of handwritten notes.

Problem 1.1

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

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

We here provide solutions for the questions of Problem 1.1 in the order they appear in the instructions.

(i) Prove that (1) is equivalent to a minimum positive curvature $\nabla^2 f(\boldsymbol{x}) \succcurlyeq \mu \boldsymbol{I},\ \forall \boldsymbol{x} \in \mathcal{X}$.

Solution: We first assume $\mathcal{X} \subset \mathbb{R}^n$. For a twice differentiable function f we have from Taylor's theorem that

$$f(\boldsymbol{x}_2) = f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{1}{2}(\boldsymbol{x}_2 - \boldsymbol{x}_1)^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}_3)(\boldsymbol{x}_2 - \boldsymbol{x}_1), \tag{2}$$

for some $\boldsymbol{x}_3$ on the interval $[\boldsymbol{x}_1,\ \boldsymbol{x}_2]$.
Combining (1) with (2) we obtain

$$\begin{aligned} f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{1}{2}(\boldsymbol{x}_2 - \boldsymbol{x}_1)^{\mathrm{T}}\nabla^2 f(\boldsymbol{x}_3)(\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \\ f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2, \end{aligned} \tag{3}$$

which is equivalent to

$$(\boldsymbol{x}_2 - \boldsymbol{x}_1)^{\mathrm{T}}(\nabla^2 f(\boldsymbol{x}_3) - \mu \boldsymbol{I})(\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq 0, \tag{4}$$

or

$$\nabla^2 f(\boldsymbol{x}) \succcurlyeq \mu \boldsymbol{I},\ \forall x \in \mathcal{X}. \tag{5}$$

We have now proved that $\nabla^2 f(\boldsymbol{x}) \succcurlyeq \mu \boldsymbol{I} \Leftrightarrow$ (1).

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

Solution: From (1) we have

$$f(\boldsymbol{x}_2) - f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2, \tag{6}$$

$$f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_2)^{\mathrm{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2) \geq \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2, \tag{7}$$

If we sum (6) and (7), we obtain

$$\mu\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2 \leq (\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1), \tag{8}$$

which concludes the proof.

(iii) Prove that (1) implies

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

Solution: (1) provides that $f(\boldsymbol{x}_2)$ is convex w.r.t. any fixed $\boldsymbol{x}_1 = \boldsymbol{x}$. We take the gradient of the righthand side of (1) w.r.t. $\boldsymbol{x}$ and set it to zero to find the minimizing $\boldsymbol{x}^*$ (minimizing the righthand side) as

$$\nabla f(\boldsymbol{x}) + \frac{\mu}{2}(2\boldsymbol{x}^* - 2\boldsymbol{x}) = 0, \tag{9}$$

$$\Rightarrow \boldsymbol{x}^* = \boldsymbol{x} - \frac{1}{\mu}\nabla f(\boldsymbol{x}). \tag{10}$$

From (1) we have

$$f^* \geq f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|_2^2 \tag{11}$$

$$\begin{aligned} &= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{x} - \frac{1}{\mu}\nabla f(\boldsymbol{x}) - \boldsymbol{x}) \\ &\quad + \frac{\mu}{2}\|\boldsymbol{x} - \frac{1}{\mu}\nabla f(\boldsymbol{x}) - \boldsymbol{x}\|_2^2 \end{aligned} \tag{12}$$

$$= f(\boldsymbol{x}) - \frac{1}{\mu}\nabla f(\boldsymbol{x})^{\mathrm{T}}\nabla f(\boldsymbol{x}) + \frac{\mu}{2}\|\frac{1}{\mu}\nabla f(\boldsymbol{x})\|_2^2 \tag{13}$$

$$= f(\boldsymbol{x}) - \frac{1}{\mu}\|\nabla f(\boldsymbol{x})\|_2^2 + \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2 \tag{14}$$

$$= f(\boldsymbol{x}) - \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2, \tag{15}$$

for any $\boldsymbol{x} \in \mathcal{X}$. Rearranging the terms concludes the proof:

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

(b) $\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \leq \frac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2,\ \forall \boldsymbol{x} \in \mathcal{X}$.

Solution: From (ii), we have $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \mu\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$. Together with Cauchy-Schwarz inequality $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} \leq \|\boldsymbol{v}\|_2\|\boldsymbol{w}\|_2$ we obtain

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

Dividing by $\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_2$ and assuming $\boldsymbol{x}_2 \neq \boldsymbol{x}_1$ so not to have $\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_2 = 0$ concludes the proof:

$$\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \leq \frac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2. \tag{18}$$

(c) $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) \leq \frac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2,\ \forall \boldsymbol{x}_1, \boldsymbol{x}_2.$

Solution: Cauchy-Schwarz together with the result in (iii) (b) gives

$$\begin{aligned}
(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) &\leq \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 &(19)\\
&\leq \|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 \frac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2 &(20)\\
&= \frac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2 &(21)
\end{aligned}$$

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

Solution: Assume r is μ-strongly convex. By the first-order condition on convexity we have for f that

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1). \tag{22}$$

By (1) we have for $r(\boldsymbol{x})$ that

$$r(\boldsymbol{x}_2) \geq r(\boldsymbol{x}_1) + \nabla r(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2. \tag{23}$$

Let $g(\boldsymbol{x})$ denote $g(\boldsymbol{x}) = f(\boldsymbol{x}) + r(\boldsymbol{x})$. It follows that $\nabla g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) + \nabla r(\boldsymbol{x})$. By the addition of (22) and (23), we have for $g(\boldsymbol{x})$ that

$$\begin{aligned}
g(\boldsymbol{x}_2) &= f(\boldsymbol{x}_2) + r(\boldsymbol{x}_2) &(24)\\
&\geq f(\boldsymbol{x}_1) + r(\boldsymbol{x}_1)\\
&\quad + \nabla f(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \nabla r(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) &(25)\\
&\quad + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2\\
&= g(\boldsymbol{x}_1) + \nabla g(\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2. &(26)
\end{aligned}$$

Which proves that $f(\boldsymbol{x}) + r(\boldsymbol{x})$ is strongly convex if f is convex and r is strongly convex.

problem 1.2

A Function $f: R^d \longrightarrow R$ is L-smooth iff it is differentiable and it's gradient is L-Lipschitz continuous (usually w.r.t norm-2):

$$\forall x_1, x_2 \in \mathbb{R}^d, \quad \|\nabla f(x_2) - \nabla f(x_1)\|_2 \leq L \|x_2 - x_1\|_2$$

For all x_1, x_2, prove that

a) $f(x_2) \leq f(x_1) + \nabla f^T(x_1)(x_2 - x_1) + \frac{L}{2}\|x_2 - x_1\|_2^2$

solution:

Let define $g(t) \triangleq f(x_1 + t(x_2 - x_1))$

we know that:

$$\int_0^1 g'(t)\, dt = g(1) - g(0) = f(x_2) - f(x_1)$$

It then follows that:

$$f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) = \int_0^1 \nabla f^T(x_1 + t(x_2 - x_1))(x_2 - x_1)\, dt - \nabla f^T(x_1)(x_2 - x_1)$$

$$\Rightarrow f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) = \int_0^1 \left(\nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1)\right)^T (x_2 - x_1)\, dt$$

from the cauchy-schartz inequality we can get:

$$\Rightarrow f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) \leq \int_0^1 \|\nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1)\|_2 \|x_2 - x_1\|_2\, dt$$

f has L-Lipschitz continuous gradient

we then have:

$$\Rightarrow f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) \leq \int_0^1 t L \|x_2 - x_1\|_2 \cdot \|x_2 - x_1\|_2 \, dt$$

$$\leq \int_0^1 t L \|x_2 - x_1\|_2^2 \, dt$$

$$\leq L \|x_2 - x_1\|_2^2 \int_0^1 t \, dt$$

$$\leq \frac{L}{2} \|x_2 - x_1\|_2^2$$

$$\Rightarrow f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) \leq \frac{L}{2} \|x_2 - x_1\|_2^2$$

$$\Rightarrow f(x_2) \leq f(x_1) + \nabla f^T(x_1)(x_2 - x_1) + \frac{L}{2} \|x_2 - x_1\|_2^2$$

b) $f(x_2) \geq f(x_1) + \nabla f^T(x_1)(x_2 - x_1) + \frac{1}{2L}\|\nabla f(x_2) - \nabla f(x_1)\|_2^2$

Let define $g_{x_1}(x_2) \triangleq f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1)$

Since f is convex therefor: $f(x_2) \geq f(x_1) + \nabla f^T(x_1)(x_2 - x_1)$

$$\Rightarrow f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) \geq 0$$

$$\Rightarrow g_{x_1}(x_2) \geq 0$$

In particular $g_{x_1}(x_1) = 0 \Rightarrow g_{x_1}(x_1) = \min_x g_{x_1}(x_2)$

and $\nabla g_{x_1}(x_1) = -\nabla f(x_1) + \nabla f(x_1) = 0$

from the optimality of x_1, it then follows that

$$g_{x_1}(x_1) \leq \min_\eta g_{x_1}(x_2 - \eta \nabla g_{x_1}(x_2))$$

$$(*) \qquad = \min_\eta f(x_2 - \eta \nabla g_{x_1}(x_2)) - f(x_1) - \nabla f^T(x_1)(x_2 - \eta \nabla g_{x_1}(x_2) - x_1)$$

By deffinition of L-smooth we have:

$$f(x_2 - \eta \nabla g_{x_1}(x_2)) \leq f(x_2) + \nabla f^T(x_2)(-\eta \nabla g_{x_1}(x_2)) + \frac{L}{2}\|\eta \nabla g_{x_1}(x_2)\|_2^2$$

In the follows from $(*)$ we have:

$$g_{x_1}(x_1) \leq \min_\eta f(x_2) + \nabla f^T(x_2)(-\eta \nabla g_{x_1}(x_2)) + \frac{L}{2}\|\eta \nabla g_{x_1}(x_2)\|_2^2$$
$$- f(x_1) - \nabla f^T(x_1)(x_2 - x_1 - \eta \nabla g_{x_1}(x_2))$$

$$g_{x_1}(x_1) \leq \min_{\eta} g_{x_1}(x_2) + \frac{L}{2}\|\eta \nabla g_{x_1}(x_2)\|_2^2 - \eta \nabla g_{x_1}^T(x_2)(\nabla f(x_2) - \nabla f(x_1))$$

$$g_{x_1}(x_1) \leq \min_{\eta} g_{x_1}(x_2) + \frac{L}{2}\eta^2 \|\nabla g_{x_1}(x_2)\|_2^2 - \eta \|\nabla g_{x_1}(x_2)\|_2^2$$

The minimum Solution η to minimize this quadratic problem is:

$$L\eta^* \|\nabla g_{x_1}(x_2)\|_2^2 - \|\nabla g_{x_1}(x_2)\|_2^2 = 0$$

$$\eta^* = \frac{1}{L}$$

$\Rightarrow$ minimum Solution : $g_{x_1}(x_2) - \frac{1}{2L}\|\nabla g_{x_1}(x_2)\|_2^2$

thus from our diffinition of $g_{x_1}(x_2)$ it follows:

$$g_{x_1}(x_1) \leq g_{x_1}(x_2) - \frac{1}{2L}\|\nabla g_{x_1}(x_2)\|_2^2$$

$$0 \leq f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) - \frac{1}{2L}\|\nabla f(x_2) - \nabla f(x_1)\|_2^2$$

$$\Rightarrow -f(x_2) \leq -f(x_1) - \nabla f^T(x_1)(x_2 - x_1) - \frac{1}{2L}\|\nabla f(x_2) - \nabla f(x_1)\|_2^2$$

$$\Rightarrow f(x_2) \geq f(x_1) + \nabla f^T(x_1)(x_2 - x_1) + \frac{1}{2L}\|\nabla f(x_2) - \nabla f(x_1)\|_2^2$$

c) $(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \frac{1}{L} \| \nabla f(x_2) - \nabla f(x_1) \|_2^2$

Let define two convex functions f_{x_1}, f_{x_2} with R^n domain

$$\begin{cases} f_{x_1}(z) = f(z) - \nabla f^T(x_1) \cdot z \\ f_{x_2}(z) = f(z) - \nabla f^T(x_2) \cdot z \end{cases}$$

This two functions have L-Lipschitz continuous gradient we know that if $f: R^n \to R$ and f has a minimizer x^* then from the inequality in problem 1.2 b we have:

$$f(z) \geq f(x^*) + \cancelto{0}{\nabla f^T(x^*)} (z - x^*) + \frac{1}{2L} \| \nabla f(z) - \cancelto{0}{\nabla f(x^*)} \|_2^2$$

$$\Rightarrow f(z) - f(x^*) \geq \frac{1}{2L} \| \nabla f(z) \|^2$$

$\Rightarrow z = x_1$ minimize $f_{x_1}(z)$

$$(1) \quad f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) = f_{x_1}(x_2) - f_{x_1}(x_1)$$
$$\geq \frac{1}{2L} \| \nabla f_{x_1}(x_2) \|_2^2$$
$$= \frac{1}{2L} \| \nabla f(x_2) - \nabla f(x_1) \|_2^2$$

similarly $z = x_2$ minimize $f_{x_2}(z)$

$$(2) \quad f(x_1) - f(x_2) - \nabla f^T(x_2)(x_1 - x_2) \geq \frac{1}{2L} \| \nabla f(x_2) - \nabla f(x_1) \|_2^2$$

now we can combine 1, 2 inequality:

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \frac{1}{L} \| \nabla f(x_2) - \nabla f(x_1) \|_2^2$$

Problem 1.3

A key performance measure of any iterative algorithm is its rate of convergence, i.e., how many iterations is needed to obtain a certain level of accuracy.

We now define the rate of convergence, or convergence rate. Let the set $\{\boldsymbol{x}_k\}$ be a sequence of updates produced by an iterative algorithm. If the algorithm converges, it will reach $\boldsymbol{x}^*$ as $k \to \infty$. Let the error $e_k = ||\boldsymbol{x}_k - \boldsymbol{x}^*||$ denote how far the current value $\boldsymbol{x}_k$ is from the optimal value $\boldsymbol{x}^*$.

To find the rate of convergence for an algorithm we investigate the following limit:

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k^p} = \lim_{k\to\infty} \frac{||\boldsymbol{x}_{k+1} - \boldsymbol{x}^*||}{||\boldsymbol{x}_k - \boldsymbol{x}^*||^p}. \tag{27}$$

Different values of this limit defines different convergence rates:

- **Sublinear:** If for $p = 1$
$$\lim_{k\to\infty} \frac{||\boldsymbol{x}_{k+1} - \boldsymbol{x}^*||}{||\boldsymbol{x}_k - \boldsymbol{x}^*||} = 1, \tag{28}$$
then the sequence $\{\boldsymbol{x}_k\}$ has sublinear convergence to $\boldsymbol{x}^*$.

- **Linear:** If for $p = 1$ and there exists some $\mu \in (0, 1)$ that fulfills,
$$\lim_{k\to\infty} \frac{||\boldsymbol{x}_{k+1} - \boldsymbol{x}^*||}{||\boldsymbol{x}_k - \boldsymbol{x}^*||} = \mu, \tag{29}$$
then the sequence $\{\boldsymbol{x}_k\}$ has linear convergence to $\boldsymbol{x}^*$.

- **Superlinear:** If for $p = 1$
$$\lim_{k\to\infty} \frac{||\boldsymbol{x}_{k+1} - \boldsymbol{x}^*||}{||\boldsymbol{x}_k - \boldsymbol{x}^*||} = 0, \tag{30}$$
then the sequence $\{\boldsymbol{x}_k\}$ has superlinear convergence to $\boldsymbol{x}^*$.

- **Quadratic:** If for $p = 2$
$$\lim_{k\to\infty} \frac{||\boldsymbol{x}_{k+1} - \boldsymbol{x}^*||}{||\boldsymbol{x}_k - \boldsymbol{x}^*||^2} < M > 0, \tag{31}$$
then the sequence $\{\boldsymbol{x}_k\}$ has quadratic convergence to $\boldsymbol{x}^*$.

Problem 1.4

Consider

$$\min_x \quad f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x_i), \tag{32}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{33}$$

for $\boldsymbol{A} \in \mathbb{R}^{p\times N}$, $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{b} \in \mathbb{R}^{p\times 1}$.

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

Solution: We can solve this using constraint free gradient descent (GD) by eliminating the equality constraint. Assuming $p < N$ (otherwise, a solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is unique), find $\boldsymbol{F}$ which spans the nullspace of $\boldsymbol{A}$, i.e., $\boldsymbol{A}\boldsymbol{F} = 0 \in \mathbb{R}^{p\times(n-1)}$. The matrix $\boldsymbol{F}$ parametrizes the feasible set as

$$\{\boldsymbol{x} : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} = \{\boldsymbol{F}\boldsymbol{z} + \hat{\boldsymbol{x}} : \ \boldsymbol{z} \in \mathbb{R}^{(n-p)\times 1}\}.$$

Choosing a particular solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, e.g., $\hat{\boldsymbol{x}} = \boldsymbol{A}^{+}\boldsymbol{b}$, where $\boldsymbol{A}^{+}$ denotes the Moore-Penrose pseudoinverse, leads to the following minimization problem:

$$\min_{z} f(\boldsymbol{F}\boldsymbol{z} + \hat{\boldsymbol{x}}). \tag{34}$$

We obtain the optimal solution as $\boldsymbol{x}^{*} = \boldsymbol{F}\boldsymbol{z}^{*} + \hat{\boldsymbol{x}}$, where $\boldsymbol{x}^{*}$ clearly is a feasible solution:

$$\boldsymbol{A}\boldsymbol{x}^{*} = \boldsymbol{A}\boldsymbol{F}\boldsymbol{z}^{*} + \boldsymbol{A}\hat{\boldsymbol{x}} = 0 + \boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{b}.$$

The optimal dual variable can be found by $\nu^{*} = -(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}\nabla f(\boldsymbol{x}^{*})$. The problem in (34) can solved using GD, by computing the gradients with respect to z which are $\nabla_{z} f(\boldsymbol{F}\boldsymbol{z} + \hat{\boldsymbol{x}}) = \boldsymbol{F}^{\mathrm{T}}\nabla f(\boldsymbol{F}\boldsymbol{z} + \hat{\boldsymbol{x}})$, where the gradient in the righthand side is with respect to the argument of f.

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

Solution: Assuming $p \ll N$, it can be more efficient to solve the Lagrange dual problem instead of (32):

$$\max_{\nu} -\boldsymbol{b}^{\mathrm{T}}\nu - f^{*}(-\boldsymbol{A}^{\mathrm{T}}\nu), \tag{35}$$

where $f^{*}(x)$ is the conjugate function to f. This is an optimization problem in p variables rather than N as in (32).

(c) Can we use Newton's method for $N = 109$? Try efficient method for computing $\nabla^{2} f(\boldsymbol{x}_{k})$ for $p = 1$ and $b = 1$ (probability simplex constraint). Extend it to $1 \leq p \ll N$.

Solution: If we approximate f by its second order Taylor polynomial, the newton step to minimize this approximation is characterized by $\Delta\boldsymbol{x}$:w

$$\begin{bmatrix} \nabla^{2} f(\boldsymbol{x})^{2} & \boldsymbol{A}^{\mathrm{T}} \\ \boldsymbol{A} & 0 \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{x} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(\boldsymbol{x}) \\ 0 \end{bmatrix} \tag{36}$$

It will be computationally infeasible to compute the inverse of the matrix on the lefthand side. We could do stochastic GD on reduced problem in (34).

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

Solution: Assuming $r(\boldsymbol{x})$ is convex, our solutions in (a), (b) and (c) still apply.

problem 1.5

In the convergence proof of GD with constant step size and strongly convex objective function proof the coercivity of the gradient:

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

Solution:

Let define $g(x) = f(x) - \frac{\mu}{2} \|x\|_2^2$

from strong convexity of $f(x)$, we get $g(x)$ is convex

based on convergence of GD with constant step size and μ-strongly convex and L-smooth, we can say that the function $f(x)$, is differentiable and it's gradient is L-Lipschitz continuous,

So we get $g(x)$ is also L-Lipschitz continuous and smooth with parameter $(L-\mu)$

Now we can apply inequality in problem 1.2c to $g(x)$: co-coercivity

$$\left(\nabla g(x) - \nabla g(y)\right)^T (x-y) \geqslant \frac{1}{L-\mu} \|\nabla g(x) - \nabla g(y)\|_2^2$$

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

$$\left(\nabla f(x) - \nabla f(y) - \mu(x-y)\right)^T (x-y) \geqslant \frac{1}{L-\mu} \|\nabla f(x) - \nabla f(y) - \mu(x-y)\|$$

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

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

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

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

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