Homework I (Jan 23^{rd}, 2019)

Group 5 (Yibei Li, Boris Petkovic, Hao Chen, Wenjun Xiong, Wenqing Yan)

HW1 (a) Prove properties of strong convexity.

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2 \quad \text{—— } (*)$$

(1) Prove that $(*)$ is equivalent to $\nabla^2 f(x) \succeq \mu I_d \quad \forall x \in \mathbb{X}$.

proof: Assume $f(x)$ is twice continuously differentiable. Then its Taylor expansion can be written as:

$$f(x_2) = f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{1}{2} (x_2 - x_1)^T \nabla^2 f(y) (x_2 - x_1)$$

for some $y \in \{\theta x_1 + (1-\theta) x_2 \mid 0 \leq \theta \leq 1\}$.

① Necessity:

If $f(x)$ is strongly convex, by $(*)$ we know that:

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

for any $x_1, x_2 \in \mathbb{X}$.

Then we have

$$(x_2 - x_1)^T (\nabla^2 f(y) - \mu I_d)(x_2 - x_1) \geq 0 \qquad \forall x_1 \in \mathbb{X},\ x_2 \in \mathbb{X}$$

Since x_1 and x_2 are arbitrary, it must holds that

$$\nabla^2 f(x) \succeq \mu I_d \qquad \forall x \in \mathbb{X}.$$

② Sufficiency:

If we have $\nabla^2 f(x) \succeq \mu I_d$, $\forall x \in \mathbb{X}$, then it holds that:

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

for any $x_1, x_2 \in \mathbb{X}$.

$\Longrightarrow f(x)$ is strongly convex.

Therefore, we have shown that $(*)$ is equivalent to $\nabla^2 f(x) \succeq \mu I_d$, $\forall x \in \mathbb{X}$.

(2) Prove that $(*)$ is equivalent to

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2 \qquad \text{—— } (*2)$$

① Necessity:

Assume $f(x)$ is strongly convex, then by $(*)$ we know that:

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

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

Taking the sum of above equations, it holds that

$$0 \geq (\nabla f(x_1) - \nabla f(x_2))^T (x_2 - x_1) + \mu \|x_2 - x_1\|_2^2$$

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

Thus, we have $(*) \Rightarrow (*2)$

② Sufficiency:

By $(*2)$ we have that:

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

$$\Rightarrow [(\nabla f(x_2) - \mu x_2) - (\nabla f(x_1) - \mu x_1)]^T (x_2 - x_1) \geq 0 \qquad \text{—— } (2\text{-}1)$$

Note that $g(x)$ is a convex function iff $(\nabla g(x_2) - \nabla g(x_1))^T (x_2 - x_1) \geq 0$.

Taking $g(x) = f(x) - \frac{\mu}{2} \|x\|^2$, Eq. (2-1) means that

$$(\nabla g(x_2) - \nabla g(x_1))^T (x_2 - x_1) \geq 0 \quad \forall x_1, x_2 \in \mathbb{X}.$$

$\Rightarrow g(x)$ is convex

$\Rightarrow g(x_2) \geq g(x_1) + \nabla g(x_1)^T (x_2 - x_1)$

i.e. $f(x_2) - \frac{\mu}{2} \|x_2\|^2 \geq f(x_1) - \frac{\mu}{2} \|x_1\|^2 + (\nabla f(x_1) - \mu x_1)^T (x_2 - x_1)$

$$\Rightarrow f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \underbrace{(\|x_1\|^2 + \|x_2\|^2 - 2x_1^T x_2)}_{\|x_1 - x_2\|^2}$$

Thus, we have $(*2) \Rightarrow (*)$

Therefore, we have shown that $(*)$ is equivalent to $(*2)$.

(3) Prove that (*) implies

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

proof. Taking minimization w.r.t x_2 on both sides of (*), we have:

$$f(x^*) \geqslant \min_{x_2 \in \mathbb{X}} \{ f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2}\|x_2 - x_1\|^2 \}$$

$$= f(x_1) - \frac{1}{2\mu}\|\nabla f(x_1)\|_2^2$$

Since x_1 is arbitrary, it holds that

$$f(x) - f^* \le \frac{1}{2\mu}\|\nabla f(x)\|_2^2 \quad \forall x \in \mathbb{X}.$$

(4) Prove that (*) implies

$$\|x_2 - x_1\|_2 \le \frac{1}{\mu}\|\nabla f(x_2) - \nabla f(x_1)\|_2 \qquad \forall x_1, x_2 \in \mathbb{X}$$

proof. We have shown in (2) that (*) $\Leftrightarrow (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geqslant \mu \|x_2 - x_1\|_2^2$.

Then by Cauchy-Schwartz inequality, it holds that

$$\|\nabla f(x_2) - \nabla f(x_1)\| \cdot \|x_2 - x_1\| \geqslant \|(\nabla f(x_2) - \nabla f(x_1))^T \cdot (x_2 - x_1)\| \geqslant \mu \|x_2 - x_1\|_2^2.$$

$$\Longrightarrow \|\nabla f(x_2) - \nabla f(x_1)\| \geqslant \mu \|x_2 - x_1\| \qquad \forall x_1, x_2 \in \mathbb{X}$$

(5) Prove that (*) implies

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \le \frac{1}{\mu}\|\nabla f(x_2) - \nabla f(x_1)\|_2^2 \quad \forall x_1, x_2 \in \mathbb{X}.$$

proof. Consider the function $g(y) = f(y) - \nabla f(x_1)^T \cdot y$

Then we have:

$$(\nabla g(y_2) - \nabla g(y_1))^T (y_2 - y_1) = (\nabla f(y_2) - \nabla f(y_1))^T (y_2 - y_1) \geqslant \mu \|y_2 - y_1\|^2$$

$\Longrightarrow g(y)$ is strongly convex w.r.t. y.

Then by (3) we know that:

$$g(x_2) - g^* \le \frac{1}{2\mu}\|\nabla g(x_2)\|^2 \quad \forall x_2 \qquad \text{— (5-1)}$$

It is obvious that

$$g^* = \min_{y \in \mathbb{X}} g(y) = g(x_1) = f(x_1) - \nabla f(x_1)^T \cdot x_1$$

Plugging into (5-1), it holds that

$$f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1) \le \frac{1}{2\mu}\|\nabla f(x_2) - \nabla f(x_1)\|^2 \quad \text{— (5-2)}$$

Interchanging x_1 and x_2 in (5-2), we have that:

$$f(x_1)-f(x_2)-\nabla f(x_2)^T(x_1-x_2) \leq \frac{1}{3\mu}\|\nabla f(x_1)-\nabla f(x_2)\|^2 \quad \text{—— (5-3)}$$

Taking the sum of (5-2) and (5-3), it holds that:

$$(\nabla f(x_2)-\nabla f(x_1))^T(x_2-x_1) \leq \frac{1}{\mu}\|\nabla f(x_2)-\nabla f(x_1)\|^2 \quad \forall x_1, x_2 \in X.$$

(6) Show that $h(x)=f(x)+r(x)$ is strongly convex for any convex f and strongly convex r.

proof. f is convex $\Leftrightarrow f(x_2) \geq f(x_1)+\nabla f(x_1)^T(x_2-x_1) \quad \forall x_1, x_2 \in X.$

r is strongly convex $\Leftrightarrow r(x_2) \geq r(x_1)+\nabla r(x_1)^T(x_2-x_1)+\frac{\mu}{2}\|x_2-x_1\|^2 \quad \forall x_1, x_2$

Taking the sum of the above two inequalities, it implies:

$$r(x_2)+f(x_2) \geq f(x_1)+r(x_1)+(\nabla f(x_1)+\nabla r(x_1))^T(x_2-x_1)+\frac{\mu}{2}\|x_2-x_1\|^2$$

$$\Rightarrow h(x_2) \geq h(x_1)+\nabla h(x_1)^T(x_2-x_1)+\frac{\mu}{2}\|x_2-x_1\|^2$$

$\Rightarrow h(x)=f(x)+r(x)$ is strongly convex.

Homework I (b) (Jan 26th, 2019) Wenqing Yan

Function: $f: R^d \to R$ is L-smooth iff it is differentiable and its gradient is L-Lipschitz-continuous (w.r.t. norm-2).

$$\forall x_1, x_2 \in R^d, \ \|\nabla f(x_2) - \nabla f(x_1)\|_2 \le L \|x_2 - x_1\|_2 \quad \cdots\cdots (1)$$

$L>0$

1(a) Prove that (1) implies

$$f(x_2) \le f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{L}{2}\|x_2 - x_1\|_2^2 \quad , \forall x_1, x_2 \quad \cdots\cdots (a)$$

Solution 1:

consider the function $g(x) = \frac{L}{2} x^T x - f(x)$

Then we can prove $g(x)$ is convex:

$$\because (\nabla g(x) - \nabla g(y))^T (x-y) = [(L \cdot x - \nabla f(x)) - (L \cdot y - \nabla f(y))]^T (x-y)$$
$$= [L \cdot (x-y) - (\nabla f(x) - \nabla f(y))]^T (x-y)$$
$$= L \cdot \|x-y\|_2^2 - (\nabla f(x) - \nabla f(y))^T (x-y)$$

Then by Cauchy-Schwarz inequality: $\ge L\|x-y\|_2^2 - \|\nabla f(x) - \nabla f(y)\| \, \|x-y\| \cdots (a\text{-}1)$

$\because$ Multiply $\|x-y\|_2$ on both sides of (1), we have:

$$\|\nabla f(x_2) - \nabla f(x_1)\|_2 \cdot \|x-y\|_2 \le L \|x_2 - x_1\|_2^2$$

Plugging into (a-1) it holds that:

$$(\nabla g(x) - \nabla g(y))^T (x-y) \ge L\|x-y\|_2^2 - L\|x-y\|_2^2 = 0$$

$\therefore$ $g(x)$ is convex

Following from the first-order condition for convexity of $g(x)$.

$$g(y) \ge g(x) + \nabla g(x)^T (y-x). \quad \forall x, y$$

$$\therefore \frac{L}{2} y^T y - f(y) \ge \frac{L}{2} x^T x - f(x) + [L \cdot x - \nabla f(x)]^T (y-x).$$

$$f(y) \le f(x) + \frac{L}{2} y^T y - \frac{L}{2} x^T x - L x^T (y-x) + \nabla f(x)^T (y-x)$$

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2$$

x, y is arbitrary, let $y = x_2$, $x = x_1$

$$\therefore f(x_2) \le f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{L}{2}\|x_2 - x_1\|^2 \quad \forall x_1, x_2.$$

(a). prove (1) implies:

$$f(x_2) \le f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{L}{2} \|x_2 - x_1\|_2^2 \quad \forall x_1, x_2$$

Solution 2

Proof. We present $f(x_2) - f(x_1)$ as an integral, apply Cauchy-Schwarz and then:

$$\|f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1)\|_2$$
$$= \left\| \int_0^1 \nabla f(x_1 + t(x_2 - x_1))^T (x_2 - x_1)\, dt - \nabla f(x_1)^T (x_2 - x_1) \right\|_2$$
$$= \left\| \int_0^1 [\nabla f(x_1 + t(x_2 - x_1))^T - \nabla f(x_1)^T] (x_2 - x_1)\, dt \right\|_2.$$
$$\le \int_0^1 \|\nabla f(x_1 + t(x_2 - x_1))^T - \nabla f(x_1)^T\|_2 \cdot \|x_2 - x_1\|_2\, dt$$

plugging (1) into:

$$\le \int_0^1 \|tL(x_2 - x_1)\|_2 \cdot \|x_2 - x_1\|_2\, dt$$
$$= \frac{L}{2} \|x_2 - x_1\|_2^2.$$

$$\therefore \|f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1)\|_2 \le \frac{L}{2} \|x_2 - x_1\|_2^2$$
$$f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1) \le \frac{L}{2} \|x_2 - x_1\|_2^2.$$

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

proof. we have proved in (a) that:

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

Taking minimization w.r.t. x_2 on both sides of (a),

if dom $f = R^n$ and f has minimizer x^* then:

$$f(x^*) \leq \inf_{x_2 \in \operatorname{dom} f} \left(f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{L}{2} \| x_2 - x_1 \|_2^2 \right)$$

$\because$ minimizer is $x_2 = x_1 - (1/L) \nabla f(x_1)$

$$\therefore = f(x_1) - \frac{1}{2L} \| \nabla f(x_1) \|_2^2$$

$\because$ x_1 is arbitrary, it holds that:

$$f(x) - f(x^*) \geq \frac{1}{2L} \| \nabla f(x) \|_2^2 \quad \cdots \quad \text{(b-1)}$$

Consider function $g(y) = f(y) - \nabla f(x_1)^T y$

$$\| \nabla g(y_2) - \nabla g(y_1) \|_2 = \| \nabla f(y_2) - \nabla f(y_1) \|_2$$

$\because$ f is L-smooth, it hold that:

$$\| \nabla g(y_2) - \nabla g(y_1) \|_2 \leq L \| y - x \|_2^2$$

$\therefore$ $g(y)$ is L-smooth

Follow (b-1), we know that:

$$g(x_2) - g(y^*) \geq \frac{1}{2L} \| \nabla g(x_2) \|^2 \quad \cdots\cdots \quad \text{(b-2)}$$

$$g(y^*) = \min_{y \in \operatorname{dom} g} g(y) = g(x_1) = f(x_1) - \nabla f(x_1)^T \cdot x_1$$

Plugging into (b-2), it holds that:

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

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

(c). (1) implies:

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

proof we have proved in (b) that:

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

x_1, x_2 is arbitrary, it holds that:

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

sum up (C-1) (C-2), we have that:

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

HW1(C):

(a) Gradient descent method, as N is small number of datasets.

(b) No, because for $N=10^9$, it needs too many iteration times.

(c) No, because computational cost for Newton is $O(n^3)$, for large $N=10^9$, we could use quasi-Newton method instead, as its computational cost is $O(n^2)$ and convergence rate is Q-superlinear.

(d) The results may be the same, (not sure).
because these methods are insensitive to the objective function.

Problem (d). If $f\colon \mathbb{R}^n \to R$ is μ-strongly convex and L-smooth, then for all $x, y \in \mathbb{R}^n$

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

Solution. Let us define a function $g(x) = f(x) - \frac{\mu}{2}||x||_2^2$. Obviously $dom(g) = dom(f)$ which is a convex set. Let us show that g is convex and $(L-\mu)$-smooth. First we deal with smoothness. It is an easy computation to show $\nabla g(x) = \nabla f(x) - \mu x$. Using L-smoothness of f we get

$$\begin{aligned}
&||\nabla g(x) - \nabla g(y)||_2^2 = \\
&= ||\nabla f(x) - \mu x - \nabla f(y)||_2^2 - 2\mu||\nabla f(x) - \mu x - \nabla f(y)||_2||x-y||_2 + \mu^2||x-y||_2^2 \\
&\leq (L-\mu)^2||x-y||_2^2,
\end{aligned}$$

so g is $(L-\mu)$-smooth.

Let us now show that g is convex, as well. We have

$$\begin{aligned}
(\nabla g(y) - \nabla g(x))^T(y-x) &= (\nabla f(y) - \mu y - \nabla f(x) + \mu x)^T(y-x) \\
&= (\nabla f(y) - \nabla f(x))^T(y-x) - \mu(y-x)^T(y-x) \\
&= (\nabla f(y) - \nabla f(x))^T(y-x) - \mu||y-x||_2^2 \\
&\geq 0,
\end{aligned}$$

where the last inequality follows from μ-strong convexity of f.

Since g is $(L-\mu)$-smooth we can make the use of the inequality proved in HW1(b) (its third part), i.e. $(\nabla g(x) - \nabla g(y))^T(x-y) \geq \frac{1}{L-\mu}||\nabla g(x) - \nabla g(y)||_2^2$. This is equivalent to

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

i.e. to

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

Notice that $\left(1 + \frac{2\mu}{L-\mu}\right) = \frac{L+\mu}{L-\mu}$. Now if we multiply both sides of the previous inequality by $\frac{L-\mu}{L+\mu}$, we get

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