Fundamentals of Machine Learning Over Networks

Group 6
HW 1

$\boxed{4\_1}$ HW1 - Group 6

if $f(y) \geqslant f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}\|y-x\|^2 \Rightarrow \nabla^2 f(x) \succeq M I_d, \ \forall x \in X$

if $g(t) \triangleq f((1-t)x + ty)$ $\quad t \in [0,1] \to g'(t) = \nabla f((1-t)x+ty)^T (y-x)$

$$g''(t) = (y-x)^T \nabla^2 f((1-t)x+ty)(y-x)$$

according to Taylor series:

$$g(t) = g(0) + g'(0)t + g''(0)\frac{t^2}{2} + g'''(t)\frac{t^3}{6}$$

$$f(\underbrace{(1-t)x+ty}_{z}) = f(x) + \nabla f(x)^T (y-x)t + (y-x)^T \nabla^2 f(x)(y-x)\frac{t^2}{2} + 0 \quad \text{I}$$

from the condition:

$$f(z) \geqslant f(x) + \nabla f(x)^T (\overbrace{z-x}^{t(y-x)}) + \frac{M}{2}\|t(y-x)\|^2$$

$$\hookrightarrow f(z) \geqslant f(x) + \nabla f(x)^T (y-x)t + t^2 \cdot \frac{M}{2} \cdot \|y-x\|^2 \quad \text{II}$$

symmetric

$$\text{I}, \text{II} \Rightarrow (y-x)^T \nabla^2 f(x)(y-x)\frac{\cancel{t^2}}{\cancel{2}} \geqslant \cancel{t^2} \cdot \frac{M}{\cancel{2}}\|y-x\|^2 \qquad A \succeq 0 \Leftrightarrow x^T A x \geqslant 0$$

$$\Rightarrow (y-x)^T \nabla^2 f(x)(y-x) - M(y-x)^T I (y-x) \geqslant 0 \longrightarrow$$

$$\Rightarrow (y-x)^T(\nabla^2 f(x) - MI)(y-x)^T \geqslant 0 \Rightarrow \nabla^2 f(x) \succeq MI$$

$\boxed{4-2}$

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

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

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

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

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

f is strong convex $\Rightarrow (*)$ is valid.

in convex we have

$(\nabla g(y) - \nabla g(x))^T (y-x) \geq 0$

$\rightarrow (\nabla f(y) - \mu y - \nabla f(x) + \mu x)^T (y-x) \geq 0$

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

$\qquad \text{''} \qquad \geq \mu\|y\|^2 - \mu y^T x - \mu x^T y + \mu\|x\|^2$

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

$\boxed{4-a}$

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

we have

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

gradient$(f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}\|y-x\|_2^2) = 0$

$\hookrightarrow \nabla f(x)^T + M(y-x) = 0$

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

$\rightarrow f^* \geq f(x) + \nabla f(x)^T (x - \frac{1}{M}\nabla f(x) - x) + \frac{M}{2}\|x - \frac{1}{M}\nabla f(x) - x\|_2^2$

$f(x) - f^* \leq \frac{1}{M}\|\nabla f^T(x)\|^2 - \frac{M}{2}\|-\frac{1}{M}\nabla f(x)\|_2^2$

$$\boxed{f(x) - f^* \leq \frac{1}{2M}\|\nabla f^T(x)\|_2^2}$$

$\boxed{4-b}$

$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{M}{2}\|y-x\|_2^2 \xrightarrow{?} \|y-x\|_2 \leq \frac{1}{M}\|\nabla f(y) - \nabla f(x)\|_2$

according to Proof 4-2 in $\boxed{P2}$ for strong convexity we have:

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

according to holder's inequality $\|fg\|_1 \leq \|f\|_p \|g\|_q \quad \frac{1}{p}+\frac{1}{q}=1$

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

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

$\Rightarrow \|\nabla f(y) - \nabla f(x)\|_2 \geq \|y-x\|_2$

$\Rightarrow \frac{1}{M}\|\nabla f(y) - \nabla f(x)\|_2 \geq$

$\boxed{4\_c}$

Prove $(\nabla f(y) - \nabla f(x))^T (y - x) \leq \frac{1}{\mu} \|\nabla f(y) - \nabla f(x)\|_2^2 \quad , \forall x, y$

$\varphi_x(z) = f(z) - \nabla f(x)^T z$

we prove that $\varphi_x(z)$ is strong convex:

1. first we prove

— we use 4_2 in $\boxed{P2}$ for strong convexity

$$(\nabla \varphi_x(z_2) - \nabla \varphi_x(z_1))^T (z_2 - z_1) \overset{?}{\geq} \mu \|z_2 - z_1\|_2^2 \quad (*)$$

$$\rightarrow (\nabla f(z_2) - \cancel{\nabla f(x)^T} - \nabla f(z_1) + \cancel{\nabla f(x)^T})(z_2 - z_1)$$

$$\rightarrow (\nabla f(z_2) - \nabla f(z_1))^T (z_2 - z_1) \geq \mu \|z_2 - z_1\|_2^2$$

$$= \underbrace{(\nabla f(z_2) - \nabla f(z_1))^T (z_2 - z_1)}_{\text{Since } f \text{ is strong convex} \Rightarrow * \text{ inequality is established.}}$$

2. Since $\varphi_x(z)$ is strong convex $\longrightarrow$ PL- inequality with $z^* = x$ can be applied. (4_a, $\boxed{P3}$)

$$\varphi_x(y) - \varphi_x(x) \leq \frac{1}{2\mu} \|\nabla \varphi_x(y)\|_2^2$$

$$\rightarrow (f(y) - \nabla f(x)^T y - f(x) - \nabla f(x)^T x) \leq \frac{1}{2\mu} \|\nabla f(y) - \nabla f(x)^T\|_2^2$$

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

3. now interchanging x and y

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

$\rightarrow$ summing 2, 3 results

$$(\nabla f(y) - \nabla f(x))^T (y - x) \leq \frac{1}{\mu} \|\nabla f(y) - \nabla f(x)\|^2 \quad ✓$$

$\boxed{4\_d}$

f is convex $\longrightarrow f(y) \geq f(x) + \nabla f(x)^T (y - x)$

r is strong convex $\xrightarrow{+} r(y) \geq r(x) + \nabla r(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2$

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

$\hookrightarrow$ Definition of convex

$\boxed{P4}$

HW1-B-a

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

we define $\underline{\underline{g}}$ as $\Rightarrow$ $g(t) = f(x_1 + t(x_2 - x_1))$, $t \in \mathbb{R}$

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

$$\int_0^1 (x_2 - x_1) \nabla f(x_1 + t(x_2 - x_1))^T dt = \underbrace{(x_2 - x_1) \nabla f(x_1)^T}_{\text{add both } \oplus \text{ \& } \ominus \text{ part}, \ \ominus \text{ is under integral}} +$$

$$(x_2 - x_1)\left[\int_0^1 [\nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1)]^T dt\right] \leq \text{ quachi-schwarz inequality}$$

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

$$\leq (x_2 - x_1) \nabla f(x_1)^T + \|x_2 - x_1\|_2 \int_0^1 \underbrace{\|\nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1)\|_2}_{L\text{-smooth} \Rightarrow \ \leq L\|x_1 + t(x_2 - x_1) - x_1\|_2} dt$$

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

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

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

HW1 (b)-(b):

For any $z \in \mathbb{R}^n$, we have

f is convex $\Rightarrow f(z) \geq f(x) + \nabla f(x)^T (z-x)$

f is L-smooth $\Rightarrow f(z) \leq f(x) + \nabla f(x)^T (z-x) + \frac{L}{2}\|z-x\|_2^2$ ← HW1(b)-(a)

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

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

$$= -\left\| \sqrt{\tfrac{L}{2}}(z-x_2) - \frac{1}{\sqrt{2L}}(\nabla f(x_1) - \nabla f(x_2)) \right\|_2^2 + \frac{1}{2L}\|\nabla f(x_1) - \nabla f(x_2)\|_2^2 + \nabla f(x_1)^T (x_2-x_1)$$

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

for $z = x_2 + \frac{1}{L}(\nabla f(x_1) - \nabla f(x_2))$

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

we use smoothness for both $f(x_1)$ & $f(x_2)$ $\Rightarrow$

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

HW1-Bc

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

we add both sides $\Rightarrow$

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

HW1(C) - (Group 6)

Consider: $\min \frac{1}{N} \sum_{i \in [N]} f_i(x_i)$

s.t. $Ax = b$

for $A \in R^{P \times N}$ and $x = [x_1, \dots, x_N]^T$.

a) Assume: strong convexity and smoothness ; $N = 1000$

To solve (a) first we form a Lagrangian (dual) function:

$$\mathcal{L}(x, \lambda) = \frac{1}{N} \sum_{i \in N} f_i(x_i) + \lambda(b - Ax) \ ;$$ where λ is Lagrangian multiplier.

Since the Problem is now unconstrained, we can use the descent methods.

$$\nabla \mathcal{L} = \frac{1}{N} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} - \lambda a_1 \\ \frac{\partial f_2}{\partial x_2} - \lambda a_2 \\ \vdots \\ \frac{\partial f_n}{\partial x_n} - \lambda a_n \end{bmatrix}$$

a) For $N = 1000$; Since the Problem dimension is still small ($N = 1000$), we can use the Gradient descent method. (GD)

b) Due to the difficulty of finding Proper Co-ordinate for high-dimensional Problems GD is not a good choice for $N = 10^?$ any more. Here, for $N = 1.^?$, we can use Newton method since the Hessian matrix is diagonal and finding the inverse is easy:

$$\nabla^2 \mathcal{L} = \begin{bmatrix} \frac{\partial^2 f_1}{\partial x_1^2} & & & \\ & \frac{\partial^2 f_2}{\partial x_2^2} & & \\ & & \ddots & \\ & & & \frac{\partial^2 f_N}{\partial x_n^2} \end{bmatrix}$$

c) for the case $P = 1, N = 1.$? → Yes we can use Newton method because the Hessian matrix is diagonal and the inverse is easy to find.

for $1 < P << N, N = 1.$ → still Newton method is acceptable since:

$$\nabla^2 \mathcal{L} = \frac{1}{N} \begin{bmatrix} \frac{\partial^2 f_1}{\partial x_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{\partial^2 f_N}{\partial x_N^2} \end{bmatrix} - \begin{bmatrix} \lambda_1 a_{11} + \dots + \lambda_p a_{p1} & & \\ & \ddots & \\ & & \lambda_1 a_{1N} + \dots + \lambda_p a_{pN} \end{bmatrix}$$

again $\nabla^2 L$ is diagonal.

d)

d-a) $\mathcal{L}(x, \lambda) = \frac{1}{N} \sum f_i(x_i) + r(x) + \lambda(b - Ax)$

$$\nabla \mathcal{L} = \frac{1}{N} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} + \frac{\partial r}{\partial x_1} - \lambda a_1 \\ \frac{\partial f_2}{\partial x_2} + \frac{\partial r}{\partial x_2} - \lambda a_2 \\ \vdots \\ \frac{\partial f_N}{\partial x_N} + \frac{\partial r}{\partial x_N} - \lambda a_n \end{bmatrix} \longrightarrow \text{GD Can Solve the Problem.}$$

d-b) $N = 1.$? $P = 1$

$$\nabla^2 L = \begin{bmatrix} \frac{\partial^2 f_1}{\partial x_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{\partial^2 f_N}{\partial x_N^2} \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 r}{\partial x_1^2} & \frac{\partial^2 r}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 r}{\partial x_1 \partial x_n} \\ \frac{\partial^2 r}{\partial x_1 \partial x_2} & \frac{\partial^2 r}{\partial x_2^2} & & \vdots \\ & & & \frac{\partial^2 r}{\partial x_n^2} \end{bmatrix}$$

d-b) cont.

Because $\nabla^2 L$ is a Hermitian and Positive Semi definite, we can use Cholesky decomposition to decompose $\nabla^2 L$ into a lower Rank matrix and ~~it~~ its Conjugate Transpose. Hence, finding the inverse is easy and Newton works well.

d-c) still it is possible to use Newton. Because, we can use Cholesky decomposition and finding the inverse of lower rank matrix is easy.

In case of $P=1$ we can have ~~bee~~ closed expression of inverse matrix and incase of $1<P\leq N$, still the ~~P~~ matrix has a good structure for finding the inverse.

HW1 $\Rightarrow$ 4) $(\nabla f(x)-\nabla f(y))^T(x-y) \geq \frac{\mu L}{\mu+L}\|x-y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x)-\nabla f(y)\|_2^2$

we define g(x) as $\Rightarrow g(x) = f(x) - \frac{\mu}{2}\|x\|^2 \longrightarrow \underline{g}$ is convex, $\boxed{\nabla g(x) = \nabla f(x) - \mu x}$

g(x) is also L-smooth $\rightarrow$ with parameter $\underline{L-\mu} \longrightarrow$

e.g. $\Rightarrow \underline{g(x)}$ is $(L-\mu)$-smooth $\Rightarrow$ [proof is in the last page]

$(\nabla g(x)-\nabla g(y))^T(x-y) \geq \frac{1}{L-\mu}\|\nabla g(x)-\nabla g(y)\|_2^2 \xrightarrow{\text{replace}}$

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

we use the result of HW1, prob. 5 $\rightarrow$

$\Rightarrow ((\nabla f(x)-\nabla f(y)) - \mu(x-y))^T((\nabla f(x)-\nabla f(y)) - \mu(x-y)) =$

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

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

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

$2\mu(x-y)(\nabla f(x)-\nabla f(y))^T\Big) \xrightarrow[\text{simple}]{\text{make it}} (\nabla f(x)-\nabla f(y))^T(x-y)\left[1+\frac{2\mu}{L-\mu}\right] \geq$

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

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

proof of smoothness of $g(x) \Rightarrow$

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

$$(\nabla g(y) - \nabla g(x)) = (\nabla f(y) - \mu y - \nabla f(x) + \mu x) \Rightarrow$$

$$[\nabla f(y) - \mu y - (\nabla f(x) - \mu x)]^T [\nabla f(y) - \mu y - (\nabla f(x) - \mu x)] =$$

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

$$L^2 \|y-x\|_2^2 + \mu^2 \|y-x\|_2^2 - \underbrace{2\mu (y-x)[\nabla f(y) - \nabla f(x)]^T}_{\leq L^2 \|y-x\|_2^2}$$

$$\Rightarrow \|y-x\|_2^2 (L^2 + \mu^2 - 2L^2) \leq \|y-x\|_2^2 (L^2 + \mu^2 - 2mL) = \underbrace{(L-\mu)^2 \|y-x\|_2^2}$$

because we had $L \geq \mu \Rightarrow -2L^2 \leq -2\mu L$

$\downarrow$

g is $L-\mu$-smooth