ELEN E6876 Sparse and Low-Dimensional Models for Hi
Homework #1

Chenye Yang cy2540@columbia.edu

February 26, 2020

P1

(1)

According to the question,

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \\ &= \frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \end{aligned}$$

Thus, gradient and hessian of f are

$$\begin{aligned} \nabla f &= -\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \\ \nabla^2 f &= \boldsymbol{A}^T\boldsymbol{A} \end{aligned}$$

Because $\text{rank}(\boldsymbol{A}) = n$, then $\boldsymbol{A}\boldsymbol{x} \neq 0$. Thus, $(\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} > 0$. Then $\nabla^2 f = \boldsymbol{A}^T\boldsymbol{A}$ is a positive definite matrix. Therefore, the solution of $\nabla f = 0$ is the global (unique) minimum point of f:

$$\nabla f = -\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) = 0$$

$\Rightarrow$

$$-\boldsymbol{A}^T\boldsymbol{y} + \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = 0$$

$\Rightarrow$

$$\boldsymbol{x} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

In summary, the unique minimizer of f is given by

$$\boldsymbol{x}_* = \boldsymbol{A}^\dagger\boldsymbol{y} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

(2)

Claim:
$\boldsymbol{B}$ is a positive definite $n \times n$ matrix and $\mu = \lambda_{\min}(\boldsymbol{B}),\ L = \lambda_{\max}(\boldsymbol{B})$, then $\forall n \times 1$ vector satisfies the inequality

$$\mu \times \|\boldsymbol{x}\| \leq \|\boldsymbol{B}\boldsymbol{x}\| \leq L \times \|\boldsymbol{x}\|$$

Proof of claim:
The matrix $\boldsymbol{B}$ can be decomposed to $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$, $\boldsymbol{Q}$ is an unit orthogonal matrix and $\boldsymbol{\Lambda}$ is the eigenvalue matrix. In the matrix multiplication $\boldsymbol{Bx}$, the rotating is controlled by $\boldsymbol{Q}$ and the stretching is controlled by $\boldsymbol{\Lambda}$. Therefore, we have the inequality about the norm of vector.

Proof of statement

$$\begin{aligned}
||\boldsymbol{x}^{k+1} - \boldsymbol{x}_*|| &= \left\|\boldsymbol{x}^k - t^k \nabla f\left(\boldsymbol{x}^k\right) - \boldsymbol{x}_*\right\| \\
&= \left\|\boldsymbol{x}^k + t^k \boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^k\right) - \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{y}\right\| \\
&= \left\|\left(\boldsymbol{I_n} - t^k\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}^k + \left(t^k - \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\right)\boldsymbol{A}^T\boldsymbol{y}\right\| \\
&= \left\|\left(\boldsymbol{I_n} - t^k\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}^k + \left(t^k\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{I_n}\right)\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{y}\right\| \\
&= \left\|\left(\boldsymbol{I_n} - t^k\boldsymbol{A}^T\boldsymbol{A}\right)\left(\boldsymbol{x}^k - \boldsymbol{x}_*\right)\right\| \\
&= \left\|\boldsymbol{x}^k - \boldsymbol{x}_*\right\| - t^k\left\|\boldsymbol{A}^T\boldsymbol{A}\left(\boldsymbol{x}^k - \boldsymbol{x}_*\right)\right\|
\end{aligned}$$

According to the claim, $\mu\left\|\left(\boldsymbol{x}^k - \boldsymbol{x}_*\right)\right\| \le \left\|\boldsymbol{A}^T\boldsymbol{A}\left(\boldsymbol{x}^k - \boldsymbol{x}_*\right)\right\| \le L\left\|\left(\boldsymbol{x}^k - \boldsymbol{x}_*\right)\right\|$, then we have

$$\begin{aligned}
\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}_*\right\| &\le \left(1 - t^k L\right)\left\|\boldsymbol{x}^k - \boldsymbol{x}_*\right\| \quad t^k < 0 \\
\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}_*\right\| &\le \left(1 - t^k \mu\right)\left\|\boldsymbol{x}^k - \boldsymbol{x}_*\right\| \quad t^k > 0
\end{aligned}$$

Therefore, the statement holds:

$$\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}_*\right\| \le \max\left\{\left|1 - t^{(k)}L\right|, \left|1 - t^{(k)}\mu\right|\right\}\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_*\right\|$$

(3)

The optimal step size t_* will make the $\boldsymbol{x}^{(k+1)}$ as close to $\boldsymbol{x}_*$ as possible, compared to $\boldsymbol{x}^{(k)}$. Namely, $\max\left\{\left|1 - t^{(k)}L\right|, \left|1 - t^{(k)}\mu\right|\right\}$ should be minimized with the optimal step size t_*. Because the function max() reserves the convexity, the optimal step size t_* should be the intersection of two absolute functions and satisfies following

$$tL - 1 = 1 - t\mu$$

$\Rightarrow$

$$t_* = \frac{2}{L + \mu}$$

Then

$$\max\left\{|1 - tL|, |1 - t\mu|\right\} = \frac{L - \mu}{L + \mu} = \frac{\kappa - 1}{\kappa + 1}$$

The distance between iteration result and the optimal solution is bounded by $\frac{\kappa-1}{\kappa+1}$, after the iteration, we can achieve geometric rate of convergence:

$$\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_*\right\| \le \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}_*\right\|$$

The statement in question holds.

P2

(1)

Let set $A = \{x : f(x) \leq \boldsymbol{M}f(x)\}$, then $P[A] \geq \frac{1}{2}$. For the 1-Lipschitz function f, we have

$$A_t = \{x \in \mathcal{X} : d(x, A) < t\} \subset \{x : f(x) < \boldsymbol{M}f(x) + t\}$$

According to the definition of concentration function $\alpha(t)$

$$\alpha(t) = \sup_{A \subset X : \mathbb{P}[A] \geq 1/2} \mathbb{P}[d(X, A) \geq t] = \sup_{A \subset X : \mathbb{P}[A] \geq 1/2} \mathbb{P}\left[A_t^c\right]$$

Then we have

$$\mathbb{P}[f \geq \boldsymbol{M}f(X) + t] \leq \alpha(t)$$

Take f as $-f$, we have

$$\mathbb{P}[f \leq \boldsymbol{M}f(X) - t] \leq \alpha(t)$$

(2)

Because $x \sim \mathcal{N}(0, 1)$, we have

$$P[x > t] = \int_t^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Let $\phi(x) = e^{-\frac{x^2}{2}}$, we have

$$\begin{aligned} P[x > t] &= \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} \frac{1}{x}\phi(x) x dx \\ &= \frac{1}{\sqrt{2\pi}} \times \left(\frac{1}{t}\phi(t) - \int_t^{+\infty} \frac{1}{x^2}\phi(x) dx\right) \end{aligned}$$

Because

$$\int_t^{+\infty} \frac{1}{x^2}\phi(x) dx \geq 0$$

Thus

$$P[x > t] \leq \frac{1}{\sqrt{2\pi}} \times \frac{1}{t}\phi(t) = \frac{1}{t}\frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$$

The part of expression of $P[x > t]$ can be written as

$$\int_t^{+\infty} \frac{1}{x^2}\phi(x) dx = \int_t^{+\infty} \frac{1}{x^3}\phi(x) x dx$$

Thus expression of $P[x > t]$ can be written as

$$P[x > t] = \frac{1}{\sqrt{2\pi}} \left(\left(\frac{1}{t} - \frac{1}{t^3}\right)\phi(t) + \int_t^{+\infty} \frac{3}{x^4}\phi(x) dx\right)$$

Because

$$\int_t^{+\infty} \frac{3}{x^4}\phi(x) dx \geq 0$$

Thus

$$P[x > t] \geq \left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}$$

In summary, the statement is proved

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \leq \mathbb{P}[x > t] \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

(3)

$$\mathbb{E}|X - c| = \int_{-\infty}^{c} P[X \leq t]dt + \int_{c}^{+\infty} P[X \geq t]dt = F(c)$$
$$F'(c) = P[X \leq c] - P[X \geq c]$$

When c is less than median, $F^{'}(c) < 0$. When c is greater than median, $F^{'}(c) > 0$. Therefore, when $c = \boldsymbol{M}(X)$, $\mathbb{E}|X - c|$ is minimized.
According to Jensen's inequality

$$\begin{aligned} |\boldsymbol{M}(Z) - \mathbb{E}(Z)| &= |\mathbb{E}(\boldsymbol{M}(Z) - Z)| \\ &\leq \mathbb{E}(|\boldsymbol{M}(Z) - Z|) \\ &\leq \mathbb{E}(|Z - \mathbb{E}(Z)|) \\ &\leq \sqrt{\mathbb{E}\left(|Z - \mathbb{E}(Z)|^2\right)} \\ &= \sqrt{\mathrm{Var}(Z)} \end{aligned}$$

The statement is proved.

(4)

$$\mathrm{Var}(\|x\|) = \mathbb{E}\left(\|x\|^2\right) - (\mathbb{E}\|x\|)^2$$

$\Rightarrow$

$$\mathbb{E}\|x\| = \sqrt{\mathbb{E}\left(\|x\|^2\right) - \mathrm{Var}\,\|x\|} = \sqrt{n - \mathrm{Var}\,\|x\|}$$

Because $0 \leq \mathrm{Var}\,\|x\| \leq 1$, thus

$$\sqrt{n-1} \leq \mathbb{E}\|x\| \leq \sqrt{n}$$

From (2) we have

$$\mathbb{P}[f(X) - \boldsymbol{M}f(X) \geq t] \leq \Phi(t) \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

From (3) we have

$$|\boldsymbol{M}Z - \mathbb{E}Z| \leq \sqrt{\mathrm{Var}(Z)}$$

$\Rightarrow$

$$|\boldsymbol{M}f(x) - \mathbb{E}f(x)| \leq \sqrt{\mathrm{Var}(f(x))}$$

$\Rightarrow$

$$-\sqrt{\mathrm{Var}(f(x))} \leq \boldsymbol{M}f(x) - \mathbb{E}f(x) \leq \sqrt{\mathrm{Var}(f(x))}$$

$\Rightarrow$

$$\mathbb{P}[f(X) - \mathbb{E}f(x) - \sqrt{\mathrm{Var}(f(x))} \geq t] \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

Because $\mathrm{Var}(f(X)) \leq 1$

$$\mathbb{P}[f(X) - \mathbb{E}f(x) - 1 \geq t] \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$