ORF 525: Statistical Foundations of Data Science

Jianqing Fan — Frederick L. Moore'18 Professor of Finance

Problem Set #2 Spring 2021

Due Friday, February 26, 2021.

1. (a) Define $F(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}\|_2^2$, $G(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$ and $H(\boldsymbol{\beta}) = F(\boldsymbol{\beta}) + G(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathbb{R}^p$; $f(\alpha) = F[\alpha\widehat{\boldsymbol{\beta}}_1 + (1-\alpha)\widehat{\boldsymbol{\beta}}_2]$ and $g(\alpha) = G[\alpha\widehat{\boldsymbol{\beta}}_1 + (1-\alpha)\widehat{\boldsymbol{\beta}}_2]$ for $\alpha \in \mathbb{R}$.

   On the one hand, $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are minimizers of a convex function H. For any $\alpha \in [0,1]$, we have

$$H(\widehat{\boldsymbol{\beta}}_1) \le H[\alpha\widehat{\boldsymbol{\beta}}_1 + (1-\alpha)\widehat{\boldsymbol{\beta}}_2] \le \alpha H(\widehat{\boldsymbol{\beta}}_1) + (1-\alpha)H(\widehat{\boldsymbol{\beta}}_2) = H(\widehat{\boldsymbol{\beta}}_1). \tag{1}$$

   On the other hand, F and G are convex in $\boldsymbol{\beta}$. Hence f, g are convex in α, and

$$F[\alpha\widehat{\boldsymbol{\beta}}_1 + (1-\alpha)\widehat{\boldsymbol{\beta}}_2] = f(\alpha) \le \alpha f(1) + (1-\alpha)f(0) = F(\widehat{\boldsymbol{\beta}}_1), \tag{2}$$

$$G[\alpha\widehat{\boldsymbol{\beta}}_1 + (1-\alpha)\widehat{\boldsymbol{\beta}}_2] = g(\alpha) \le \alpha g(1) + (1-\alpha)g(0) = G(\widehat{\boldsymbol{\beta}}_1). \tag{3}$$

   From (1) and $H = F + G$ we see that the equalities must be achieved in both (2) and (3). Thus $f(\alpha) = F[\alpha\widehat{\boldsymbol{\beta}}_1 + (1-\alpha)\widehat{\boldsymbol{\beta}}_2] = F(\widehat{\boldsymbol{\beta}}_1)$ holds for all $\alpha \in [0,1]$, forcing $f' = 0$ over $(0,1)$.

   Observe that

$$f(\alpha) = \frac{1}{2n}\|\mathbf{X}[\alpha\widehat{\boldsymbol{\beta}}_1 + (1-\alpha)\widehat{\boldsymbol{\beta}}_2] - \mathbf{Y}\|_2^2 = \frac{1}{2n}\|\alpha\mathbf{X}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2) + \mathbf{X}\widehat{\boldsymbol{\beta}}_2 - \mathbf{Y}\|_2^2,$$

   and

$$f'(\alpha) = \frac{1}{n}[\mathbf{X}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)]^\top[\alpha\mathbf{X}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2) + \mathbf{X}\widehat{\boldsymbol{\beta}}_2 - \mathbf{Y}] = \frac{\alpha}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)\|_2^2 + \frac{1}{n}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)^\top\mathbf{X}^\top(\mathbf{X}\widehat{\boldsymbol{\beta}}_2 - \mathbf{Y})$$

   is a linear function of α, which we have shown to be zero on $(0,1)$. This finally leads to $\|\mathbf{X}(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2)\|_2^2 = 0$ and $\mathbf{X}\widehat{\boldsymbol{\beta}}_1 = \mathbf{X}\widehat{\boldsymbol{\beta}}_2$.

   (b) The first-order optimality condition for $\widehat{\boldsymbol{\beta}}$ reads $\mathbf{0} \in \partial H(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\mathbf{X}^\top(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Y}) + \lambda\partial\|\widehat{\boldsymbol{\beta}}\|_1$, where $(\partial\|\boldsymbol{\beta}\|_1)_j = \begin{cases} \{1\}, & \text{if } \beta_j > 0 \\ \{-1\}, & \text{if } \beta_j < 0 \\ [-1,1], & \text{if } \beta_j = 0 \end{cases}$. Then the desired result directly follows.

   (c) Note that by the condition $\lambda > \|n^{-1}\mathbf{X}\mathbf{y}\|_\infty$, $\boldsymbol{\beta} = 0$ is a sufficient condition of Theorem 2.1. Therefore, it is the minimizer.

2. (a) When $\lambda_2 > 0$, the loss function is strongly convex. Consequently, the minimizer $\widehat{\boldsymbol{\beta}}$ is unique.

   (b) Since the function is strongly convex, translating the first order-condition in Theorem 2.1, we have

$$n^{-1}\mathbf{X}_1^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + \lambda_1 \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_1) + 2\lambda_2\widehat{\boldsymbol{\beta}}_1 = \mathbf{0},$$
$$\|n^{-1}\mathbf{X}_2^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\|_\infty \le \lambda_1,$$

(c) A quick solution is that $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ satisfies the above condition and hence is the unique minimizer. An alternative solution is as follows. Recall from Problem 1 (c) that when $\lambda_1 > \|n^{-1}\mathbf{X}^\top \mathbf{Y}\|_\infty$, the loss function $\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1$ has the minimizer $\widehat{\boldsymbol{\beta}}_{\lambda_1} = \mathbf{0}$. Since $\mathbf{0}$ is also the unique minimizer of $\lambda_2\|\boldsymbol{\beta}\|_2^2$, one arrives at the conclusion that $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ is the unique minimizer of $\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2$.

3. (a) Fix any $\mathbf{a} \in \mathbb{R}^n$ with $\|\mathbf{a}\|_2 = 1$ and let $Z = \mathbf{a}^T\boldsymbol{\varepsilon}$. Since $\boldsymbol{\varepsilon}$ is σ-sub-Gaussian, we have

$$\mathbb{E}e^{tZ} = \mathbb{E}e^{(t\mathbf{a})^T\boldsymbol{\varepsilon}} \le \mathbb{E}e^{\|t\mathbf{a}\|_2^2\sigma^2/2} \le \mathbb{E}e^{t^2\sigma^2/2}, \qquad \forall t \in \mathbb{R}.$$

On the other hand, we obtain from Taylor expansion that $\mathbb{E}e^{t^2\sigma^2/2} = 1 + \frac{\sigma^2}{2}t^2 + o(t^2)$ and $\mathbb{E}e^{tZ} = 1 + t(\mathbb{E}Z) + \frac{t^2}{2}\mathbb{E}Z^2 + o(t^2)$ as $t \to 0$. Comparing the coefficients gives $0 = \mathbb{E}Z = \mathbf{a}^T\mathbb{E}\boldsymbol{\varepsilon}$ and $\sigma^2 \ge \mathbb{E}Z^2 = \mathbf{a}^T\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\mathbf{a}$. This finishes the proof as $\mathbf{a}$ is an arbitrary unit-norm vector.

(b) Note that $\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty = \max_{1\le j\le p}|\mathbf{X}_j^T\boldsymbol{\varepsilon}|$ and $\mathbb{E}e^{t\mathbf{X}_j^T\boldsymbol{\varepsilon}} \le \mathbb{E}e^{\|t\mathbf{X}_j\|_2^2\sigma^2/2} = \mathbb{E}e^{t^2n\sigma^2/2}$, $\forall t$. By Chebychev's inequality, for any $x \in \mathbb{R}$ and $t > 0$ we have

$$\mathbb{P}(\mathbf{X}_j^T\varepsilon > x) = \mathbb{P}(e^{t\mathbf{X}_j^T\varepsilon} > e^{tx}) = \mathbb{E}e^{t\mathbf{X}_j^T\boldsymbol{\varepsilon}}/e^{tx} \le \mathbb{E}e^{(t^2n\sigma^2/2)-tx}.$$

Taking $t = x/(n\sigma^2)$ yields $\mathbb{P}(\mathbf{X}_j^T\varepsilon > x) \le e^{-x^2/(2n\sigma^2)}$. We can also show $\mathbb{P}(\mathbf{X}_j^T\varepsilon < -x) \le e^{-x^2/(2n\sigma^2)}$ in a similar way. Then $\mathbb{P}(|\mathbf{X}_j^T\varepsilon| > x) \le 2e^{-x^2/(2n\sigma^2)}$ holds for any x and j. Therefore,

$$\mathbb{P}\left(\|n^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty > \sqrt{2(1+\delta)}\sigma\sqrt{\frac{\log p}{n}}\right) \le \sum_{j=1}^{p}\mathbb{P}\left(|\mathbf{X}_j^T\boldsymbol{\varepsilon}| > \sqrt{2(1+\delta)}\sigma\sqrt{n\log p}\right)$$
$$\le p\cdot 2\exp\left[-\frac{1}{2n\sigma^2}\left(\sqrt{2(1+\delta)}\sigma\sqrt{n\log p}\right)^2\right] = 2p^{-\delta}.$$

4. (a) By Taylor's theorem, one has

$$f(\mathbf{x}) = f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^\top(\mathbf{x} - \mathbf{x}_{i-1}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_{i-1})^\top f''(\tilde{\mathbf{x}})(\mathbf{x} - \mathbf{x}_{i-1})$$

for some $\tilde{\mathbf{x}}$ which lies between $\mathbf{x}$ and $\mathbf{x}_{i-1}$. Since we know $f''(\tilde{\mathbf{x}}) \le L$, we further get

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^\top(\mathbf{x} - \mathbf{x}_{i-1}) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_{i-1}\|_2^2 \\ &\le f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^\top(\mathbf{x} - \mathbf{x}_{i-1}) + \frac{1}{2\delta}\|\mathbf{x} - \mathbf{x}_{i-1}\|_2^2 \qquad (4) \end{aligned}$$

where the last line arises from the assumption that $\delta \le 1/L$.

(b) Taking gradient of the upper bound with respect to $\mathbf{x}$ yields

$$f'(\mathbf{x}_{i-1}) + \frac{1}{\delta}(\mathbf{x} - \mathbf{x}_{i-1}).$$

Setting this equal to zero results in the solution

$$\mathbf{x}_i = \mathbf{x}_{i-1} - \delta f'(\mathbf{x}_{i-1}).$$

(c) Apply (4) with $\mathbf{x} = \mathbf{x}_i$ to get

$$\begin{aligned} f(\mathbf{x}_i) &\le f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^\top (\mathbf{x}_i - \mathbf{x}_{i-1}) + \frac{1}{2\delta}\|\mathbf{x}_i - \mathbf{x}_{i-1}\|_2^2 \\ &= f(\mathbf{x}_{i-1}) - \frac{1}{2\delta}\|\mathbf{x}_i - \mathbf{x}_{i-1}\|_2^2, \end{aligned}$$

where the equality follows from the update rule $\mathbf{x}_i = \mathbf{x}_{i-1} - \delta f'(\mathbf{x}_{i-1})$. An immediate consequence is that

$$f(\mathbf{x}_i) \le f(\mathbf{x}_{i-1}). \tag{5}$$

In addition, utilizing the convexity, one has

$$f(\mathbf{x}^\star) \ge f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^\top (\mathbf{x}^\star - \mathbf{x}_{i-1}),$$

which implies

$$\begin{aligned} f(\mathbf{x}_i) &\le f(\mathbf{x}^\star) - f'(\mathbf{x}_{i-1})^\top (\mathbf{x}^\star - \mathbf{x}_{i-1}) - \frac{1}{2\delta}\|\mathbf{x}_i - \mathbf{x}_{i-1}\|_2^2 \\ &= f(\mathbf{x}^\star) + \frac{1}{\delta}(\mathbf{x}_{i-1} - \mathbf{x}_i)^\top (\mathbf{x}_{i-1} - \mathbf{x}^\star) - \frac{1}{2\delta}\|\mathbf{x}_i - \mathbf{x}_{i-1}\|_2^2 \\ &= f(\mathbf{x}^\star) + \frac{1}{\delta}(\mathbf{x}_{i-1} - \mathbf{x}^\star + \mathbf{x}^\star - \mathbf{x}_i)^\top (\mathbf{x}_{i-1} - \mathbf{x}^\star) - \frac{1}{2\delta}\|\mathbf{x}_i - \mathbf{x}^\star + \mathbf{x}^\star - \mathbf{x}_{i-1}\|_2^2 \\ &= f(\mathbf{x}^\star) + \frac{1}{2\delta}\left(\|\mathbf{x}_{i-1} - \mathbf{x}^\star\|_2^2 - \|\mathbf{x}_i - \mathbf{x}^\star\|_2^2\right). \end{aligned}$$

(d) Summing both hands up from $i = 1$ to k gives

$$\begin{aligned} \sum_{i=1}^k f(\mathbf{x}_i) &\le kf(\mathbf{x}^\star) + \frac{1}{2\delta}\sum_{i=1}^k \left(\|\mathbf{x}_{i-1} - \mathbf{x}^\star\|_2^2 - \|\mathbf{x}_i - \mathbf{x}^\star\|_2^2\right) \\ &= kf(\mathbf{x}^\star) + \frac{1}{2\delta}\left(\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2 - \|\mathbf{x}_k - \mathbf{x}^\star\|_2^2\right) \\ &\le kf(\mathbf{x}^\star) + \frac{1}{2\delta}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2. \end{aligned}$$

Note from (5) that $f(\mathbf{x}_i)$ is non-increasing. This further implies

$$f(\mathbf{x}_k) \le \frac{1}{k}\sum_{i=1}^k f(\mathbf{x}_i) \le f(\mathbf{x}^\star) + \frac{1}{2\delta k}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2,$$

which completes the proof.

5. cf R code.

6. cf R code.