CMSE 820: Homework #2

Due on September 21, 2019 at 11:59pm

Professor Yuying Xie

Boyao Zhu

Problem 1

Solution
Properties of Norm
Let V be a vector space, $\|\cdot\| : V \to \mathbb{R}$ is a norm:

- $\forall v \in V : \|v\| \geq 0$ and $\|v\| = 0 \iff v = 0$ (positive/definite)
- $\forall v \in V, \lambda \in \mathbb{R} : |\lambda|\|v\| = \|\lambda v\|$ (absolutely scaleable)
- $\forall v, w \in V : \|v + w\| \leq \|v\| + \|w\|$ (Triangle inequality)

For any give norm $\|\cdot\|$ and radius $r > 0$, the norm ball $B(r) = \{x : \|x\| \leq r, x \in \mathbb{R}^p\}$. $\forall x, y \in B(r), \forall t \in [0, 1]$, by the triangle inequality and the absolutely scaleable for norms, we have

$$\|tx + (1 - t)y\| \leq t\|x\| + (1 - t)\|y\| \leq r$$

It follows that $tx + (1 - t)y \in B(r)$, and hence $B(r)$ is a convex set.

Problem 2

Solution
Suppose there were two optimal solutions $x, y \in \mathbb{R}^n$. This means that $x, y \in \Omega$ and

$$f(x) = f(y) \leq f(z), \forall z \in \Omega$$

But consider $z = \frac{x+y}{2}$. By convexity of Ω, we have $z \in \Omega$. By strict convexity, we have

$$\begin{aligned} f(z) &= f\left(\frac{x + y}{2}\right) \\ &\leq \frac{1}{2}f(x) + \frac{1}{2}f(y) \\ &= \frac{1}{f}(x) + \frac{1}{2}f(x) \\ &= f(x) \end{aligned}$$

This contradicts.

Problem 3

Solution
$f(x) = |x|$. We want to show that $\partial f(0) = [-1, 1]$. It is trivial to see that

$$\forall x \in \mathbb{R}, \forall c \in [-1, 1], f(x) - cx = \begin{cases} (1 - c)x \geq 0, x \geq 0 \\ -(1 + c)x \geq 0, x < 0 \end{cases}$$

So $[-1, 1] \subseteq \partial f(0)$. Now consider any $c' > 1$. Fixing $x = 1$, we see that $f(1) < c' * 1$, so $(1, \infty) \cap \partial f(0) = \emptyset$. To sum up, $\partial f(0) = [-1, 1]$.

Problem 4

Solution
$g(x)$ is affine $\Longrightarrow \exists$ a linear function $L(x)$ such that $g(x) = L(x) + b$, where $b = g(0)$.
$\forall x, y \in \mathbb{R}^n$, $\forall t \in [-1, 1]$, we have

$$\begin{aligned} g(tx + (1-t)y) &= L(tx + (1-t)y) + b \\ &= tL(x) + (1-t)L(y) + tb + (1-t)b \\ &= t[L(x) + b] + (1-t)[L(y) + b] \\ &= tg(x) + (1-t)g(y) \\ &= 0 \end{aligned}$$

which implies that $tx + (1-t)y \in A$, and hence A is convex.

Problem 5

Solution
5.1
Let $\tilde{\mathbf{y}} = [\mathbf{y}^T \mid \mathbf{0}^T] \in \mathbb{R}^{n+p}$ and $\tilde{\mathbf{X}} = [\mathbf{X} \mid \sqrt{\lambda}\mathbf{I}] \in \mathbb{R}^{p\times(n+p)}$.

$$\tilde{y} - \tilde{X}^T\beta = \begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{X}^T\beta \\ \sqrt{\lambda}\mathbf{I}\beta\end{bmatrix} = \begin{bmatrix}\mathbf{y} - \mathbf{X}^T\beta \\ -\sqrt{\lambda}\mathbf{I}\beta\end{bmatrix}$$

$$\begin{aligned} \hat{\beta}^{ols} &= \arg\min \|\tilde{y} - \tilde{X}^T\beta\|^2 \\ &= \arg\min\{\|\tilde{y} - tildeX^T\beta\|^2 + \lambda\|\beta\|^2\} \\ &= \hat{\beta}^{ridge} \end{aligned}$$

5.2

$$\tilde{X}^T v = \begin{bmatrix}\mathbf{X}^T \\ \sqrt{\lambda}\mathbf{I}v\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

Note that since $\sqrt{\lambda}\mathbf{I}$ is full rank and $\sqrt{\lambda}\mathbf{I}v = 0$, v must be the zero vector. It follows immediately that $\tilde{\mathbf{X}}$ is full row-rank and $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ is invertible, which implies that $\hat{\beta}^{ols}$ corresponding to the predictor matrix $\tilde{X}$ is unique. According to (5.1), Ridge regression $\hat{\beta}^{ridge}$ is unique.

5.3
An explicit formula for β^{ridge}:

$$\beta^{ridge} = (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)^{-1}\tilde{\mathbf{X}}\tilde{y} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y}$$

Note that $\hat{\beta}^{ridge}$ is linear in $\mathbf{y}$. Therefore, $\forall a^T \in \mathbb{R}^p$, $a^T\hat{\beta}^{ridge}$ is also a linear function of $\mathbf{y}$

5.4
For any estimator $\hat{\theta}$, $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta})$. Since OLS estimators have the smallest variance among all linear unbiased estimators, and that Ridge estimators are also linear, it must be true that Ridge estimators are biased in order to have a smaller MSE than OLS ones. That is, $a^T\hat{\beta}^{ridge}$ is biased.

Problem 5 continued on next page...

5.5

$$\begin{aligned}\hat{\beta}^{ridge} &= (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y} \\ &= (UDV^TVDU^T + \lambda\mathbf{I})^{-1}UDV^T\mathbf{y} \\ &= (UD^2U^T + \lambda\mathbf{I})^{-1}UDV^T\mathbf{y} \\ &= (U(D^2 + \lambda\mathbf{I})U^T)^{-1}UDV^T\mathbf{y} \\ &= U(D^2 + \lambda\mathbf{I})^{-1}U^TUDV^T\mathbf{y} \\ &= U(D^2 + \lambda\mathbf{i})^{-1}DV^T\mathbf{y} \\ &= U\Lambda V^T\mathbf{y}\end{aligned}$$

where $\Lambda = \text{diag}\Big\{\frac{d_1}{d_1^2+\lambda}, \frac{d_2}{d_2^2+\lambda}, \cdots, \frac{d_r}{d_r^2+\lambda}\Big\}$.

5.6

$$\begin{aligned}\mathbb{E}(\hat{\beta}^{ridge}) &= (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbb{E}[\mathbf{y}] \\ &= (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{X}^T\beta^* \\ &= U(D^2 + \lambda\mathbf{I})^{-1}DV^TVDU^T\beta^* \\ &= U\tilde{D}U^T\beta^*\end{aligned}$$

where $\tilde{D} = \text{diag}\Big\{\frac{d_1^2}{d_1^2+\lambda}, \frac{d_2^2}{d_2^2+\lambda}, \cdots, \frac{d_r^2}{d_r^2+\lambda}\Big\}$..

$$\begin{aligned}\|\mathbb{E}(\hat{\beta}^{ridge})\| &= \|U\tilde{D}U^T\beta^*\| \\ &= \|\tilde{D}\|\|\beta^*\| \\ &< \|\beta^*\| \Longrightarrow \mathbb{E}(\hat{\beta}^{ridge}) \neq \beta^*.\end{aligned}$$

where unitarity preserves norm, and 2-norm of a matrix equals its largest singular value. That is $\|\tilde{D}\|_2 \leq 1$.