CMSE 820 HW 2

Hao Lin

October 2, 2018

1 Question 1

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 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.

2 Question 2

Let f be a strictly convex function and $f(x_0)$ be a global minimum. (Note that for convex functions, all local minima are global.)

Proof by contradiction:

If $\exists\, y \neq x_0$ such that $f(y) = f(x_0)$ is also a global minimum, then, by the strict convexity, we have

$$f(\frac{x_0 + y}{2}) < \frac{1}{2}(f(x_0) + f(y)) = f(x_0) = f(y),$$

which contradicts that $f(x_0)$ and $f(y)$ are the global minima.

Hence, $f(x_0)$ is the unique global minimum. That is, the optimal solution is unique.

3 Question 3

$f(x) = |x|$. 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 $f(1) < c' \cdot 1$, so $(1, \infty) \cap \partial f(0) = \emptyset$. Similarly, for $c' < -1$, $f(-1) < c' \cdot (-1)$, so $(-\infty, -1) \cap \partial f(0) = \emptyset$.

To sum up, $\partial f(0) = [-1, 1]$.

4 Question 4

$A = \{x \in \mathbb{R}^n : g(x) = 0\}$, where $g(x)$ is affine.
$g(x)$ is affine $\implies \exists$ 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.

5 Quesstion 5

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \{\|\mathbf{y} - \mathbf{X}^T\beta\|^2 + \lambda\|\beta\|^2\}.$$

5.1 a

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

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

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

5.2 b

Consider the following equation for some $v \in \mathbb{R}^p$.

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

In particular, since $\sqrt{\lambda}\mathbf{I}$ is full rank and $\sqrt{\lambda}\mathbf{I}v = \mathbf{0}$, v must be the zero vector. It follows immediately that, regardless of X, $\tilde{\mathbf{X}}$ is full row-rank.

Now that $\tilde{\mathbf{X}}$ is full row-rank, $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ is invertible, which implies that $\hat{\beta}^{\text{ols}}$ corresponding to the predictor matrix $\tilde{\mathbf{X}}$ is unique. According to the equivalence in (a), the corresponding Ridge regression $\hat{\beta}^{\text{ridge}}$ is also unique.

5.3 c

An explicit formula for $\hat{\beta}^{\text{ridge}}$:

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

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

5.4 d

Recall that for any estimator $\hat{\theta}$, $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})$.
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, $a^T\hat{\beta}^{\text{ridge}}$ is biased.

5.5 e

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

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

5.6 f

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

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

$\forall \lambda > 0$, the determinant $\text{Det}(\tilde{D}) = \Pi_{i=1}^{r} \frac{d_i^2}{d_i^2+\lambda} < 1$, implying that $\text{Det}(U\tilde{D}U^T) = \text{Det}(U)\text{Det}(\tilde{D})\text{Det}(U^T) < 1$. Hence,

$$\|\mathbb{E}(\hat{\beta}^{\text{ridge}})\| < \|\beta^*\| \implies \mathbb{E}(\hat{\beta}^{\text{ridge}}) \neq \beta^*.$$