1. (a) Notice that for orthogonal design matrix $\mathbf{X}$, we have $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \frac{1}{n}\mathbf{X}^\top\mathbf{Y}$. Hence,

$$\begin{aligned}\frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 &= \frac{1}{2n}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\\ &= \frac{1}{2n}(\mathbf{Y}^\top\mathbf{Y}-2\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta})\\ &= \frac{1}{2n}\mathbf{Y}^\top\mathbf{Y}-\hat{\boldsymbol{\beta}}^\top\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{p}(\hat{\beta}_j-\beta_j)^2 &= \frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\\ &= \frac{1}{2}\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^\top\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta} = \frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+h(\mathbf{X},\mathbf{Y}),\end{aligned}$$

where $h(\mathbf{X},\mathbf{Y})$ does not depend on $\boldsymbol{\beta}$. Therefore, minimizing $Q(\boldsymbol{\beta})$ is equivalent to minimizing $Q(\boldsymbol{\beta})+h(\mathbf{X},\mathbf{Y}) = \sum_{j=1}^p[\frac{1}{2}(\hat{\beta}_j-\beta_j)^2+p_\lambda(|\beta_j|)]$.

(b) From the results in part (a), we know that minimizing $Q(\boldsymbol{\beta})$ is equivalent to minimizing the sum $\sum_{j=1}^p[\frac{1}{2}(\hat{\beta}_j-\beta_j)^2+p_\lambda(\beta_j)]$. Since it is the sum of functions of each variable, the problem is further equivalent to minimize each summand $g_j(\beta_j) \equiv g_j(\beta_j|\hat{\beta}_j) = \frac{1}{2}(\hat{\beta}_j-\beta_j)^2+p_\lambda(\beta_j)$ for each j separately.

For each j with $\hat{\beta}_j \geq 0$, notice that any minimizer $\beta_j^* = \arg\min_{\beta_j} g_j(\beta_j)$ must satisfy $\beta_j^* \geq 0$ (otherwise, if $\beta_j^* < 0$, then $g_j(-\beta_j^*) = \frac{1}{2}(\hat{\beta}_j+\beta_j^*)^2+p_\lambda(-\beta_j^*) < \frac{1}{2}(\hat{\beta}_j-\beta_j^*)^2+p_\lambda(\beta_j^*) = g_j(\beta_j^*)$, resulting in a contradiction). Hence, we only need to minimize $g_j(\beta_j)$ on $\beta_j \geq 0$, i.e. it remains to solve for

$$\min_{\beta_j\geq 0} g_j(\beta_j) = \frac{1}{2}(\hat{\beta}_j-\beta_j)^2+\begin{cases}\lambda\beta_j, & 0\leq\beta_j\leq\lambda\\ \dfrac{2\gamma\lambda\beta_j-\beta_j^2-\lambda^2}{2(\gamma-1)}, & \lambda<\beta_j\leq\gamma\lambda\\ \dfrac{\lambda^2(\gamma+1)}{2}, & \beta_j>\gamma\lambda\end{cases}.$$

Differentiate the objective yields

$$g_j'(\beta_j) = \beta_j-\hat{\beta}_j+\begin{cases}\lambda, & 0<\beta_j\leq\lambda\\ \dfrac{2\gamma\lambda-2\beta_j}{2(\gamma-1)}, & \lambda<\beta_j\leq\gamma\lambda\\ 0, & \beta_j>\gamma\lambda\end{cases},$$

as we can see that both g_j and g_j' are continuous. To locate the global minimum, we split the analyses of g_j' into the following cases based on the value of $\hat{\beta}_j$:

- If $0 \leq \hat{\beta}_j \leq \lambda$, then $g_j'(\beta_j) > 0$ for all $\beta_j > 0$ (elaborate!). Hence, $g_j(\beta_j)$ is minimized iff $\beta_j = 0$.
- If $\lambda < \hat{\beta}_j \leq 2\lambda$, then $g_j'(\beta_j) = / < / > 0$ for $\beta_j = / < / > \hat{\beta}_j - \lambda$ (elaborate!). Hence, $g_j(\beta_j)$ is minimized iff $\beta_j = \hat{\beta}_j - \lambda$.
- If $2\lambda < \hat{\beta}_j \leq \gamma\lambda$, then $g_j'(\beta_j) = / < / > 0$ for $\beta_j = / < / > \frac{(\gamma-1)\hat{\beta}_j-\gamma\lambda}{\gamma-2}$ (elaborate!). Hence, $g_j(\beta_j)$ is minimized iff $\beta_j = \frac{(\gamma-1)\hat{\beta}_j-\gamma\lambda}{\gamma-2}$.

- If $\hat{\beta}_j > \gamma\lambda$, then $g_j'(\beta_j) = / < / > 0$ for $\beta_j = / < / > \hat{\beta}_j$ (elaborate!). Hence, $g_j(\beta_j)$ is minimized iff $\beta_j = \hat{\beta}_j$.

For each j with $\hat{\beta}_j \leq 0$, notice that $g_j(\beta_j|\hat{\beta}_j) = g_j(-\beta_j| - \hat{\beta}_j)$, where the minimizers of the R.H.S. are readily available since $-\hat{\beta}_j \geq 0$, i.e. we can apply the previous conclusions to $-\beta_j$, that is,

- If $0 \leq -\hat{\beta}_j \leq \lambda$, then $g_j(-\beta_j| - \hat{\beta}_j)$ is minimized iff $-\beta_j = 0$.
- If $\lambda < -\hat{\beta}_j \leq 2\lambda$, then $g_j(-\beta_j| - \hat{\beta}_j)$ is minimized iff $-\beta_j = -\hat{\beta}_j - \lambda$.
- If $2\lambda < -\hat{\beta}_j \leq \gamma\lambda$, then $g_j(-\beta_j| - \hat{\beta}_j)$ is minimized iff $-\beta_j = \frac{(\gamma-1)(-\hat{\beta}_j)-\gamma\lambda}{\gamma-2}$.
- If $-\hat{\beta}_j > \gamma\lambda$, then $g_j(-\beta_j| - \hat{\beta}_j)$ is minimized iff $-\beta_j = -\hat{\beta}_j$.

Combining the cases for $\hat{\beta}_j \geq 0$ and $\hat{\beta}_j \leq 0$, we can characterize the SCAD solutions as follows:

- If $0 \leq |\hat{\beta}_j| \leq \lambda$, then $\hat{\beta}_j^{\text{SCAD}} = 0$.
- If $\lambda < |\hat{\beta}_j| \leq 2\lambda$, then $\hat{\beta}_j^{\text{SCAD}} = \operatorname{sgn}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)$.
- If $2\lambda < |\hat{\beta}_j| \leq \gamma\lambda$, then $\hat{\beta}_j^{\text{SCAD}} = \operatorname{sgn}(\hat{\beta}_j)\frac{(\gamma-1)|\hat{\beta}_j|-\gamma\lambda}{\gamma-2}$.
- If $|\hat{\beta}_j| > \gamma\lambda$, then $\hat{\beta}_j^{\text{SCAD}} = \hat{\beta}_j$.

This is exactly what we need to prove.

2. Notice that $g(\beta)$ is continuous on $\mathbb{R}$ and continuously differentiable on $\mathbb{R}\backslash\{0\}$.

   (a) Let β_0 denote the peak of $-\beta - p_\lambda'(\beta)$ on $\beta \geq 0$. From unimodality of $-\beta - p_\lambda'(\beta)$, it follows that $t_0 \leq \beta + p_\lambda'(\beta)$ for any $\beta > 0$, and the inequality is strict except when $\beta = \beta_0$. Thus, for any $\beta > 0$,

$$g'(\beta) = \beta - \hat{\beta} + p_\lambda'(\beta) \geq \beta + p_\lambda'(\beta) - |\hat{\beta}| \geq t_0 - |\hat{\beta}| \geq 0,$$

   and the inequality is strict except when $\beta = \beta_0$. Hence g is strictly increasing on $(0, \infty)$.

   Similarly, for any $\beta < 0$,

$$g'(\beta) = \beta - \hat{\beta} - p_\lambda'(-\beta) = -\hat{\beta} - (-\beta + p_\lambda'(-\beta)) \leq -\hat{\beta} - t_0 \leq |\hat{\beta}| - t_0 \leq 0,$$

   and the inequality is strict except when $\beta = -\beta_0$, i.e. g is strictly decreasing on $(-\infty, 0)$. Hence, the unique minimizer of $g(\beta)$ is at $\beta = 0$.

   (b) The problem statement should be $|\hat{\beta}| \geq t_1$ instead of $|\hat{\beta}| \leq t_1$ for large t_1.

   Using the arguments in Problem 1(b), we know that $\hat{\beta}_\lambda$ and $\hat{\beta}$ should not have opposite signs. If $\hat{\beta} > t_1$, then $\hat{\beta}_\lambda = \arg\min_{\beta \geq 0} g(\beta)$. Since for large t_1,

$$g'(\beta) = \beta - \hat{\beta} + p_\lambda'(\beta) = / < / > 0$$

   for $\beta = / < / > \hat{\beta}$ (elaborate!), it follows that $\arg\min_{\beta \geq 0} g(\beta) = \hat{\beta}$, as desired.

   The arguments for $\hat{\beta} < -t_1$ shall be *Mutatis Mutandis*.

3. Recall from Page 24 of Lecture Note 8 that the prediction on a new observation $\mathbf{x}_0$ based on the ridge estimator is

$$\hat{\mathbf{Y}} = \mathbf{x}_0\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \sigma^2 \cdot 2^D\mathbf{I})^{-1}\mathbf{Y} = 2^{-D}\mathbf{x}_0\mathbf{X}^\top(2^{-D}\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{Y},$$

where the design matrix $\mathbf{X}$ and $\mathbf{x}_0$ are both defined as the basis functions evaluated at their original observations, i.e. $\mathbf{x}_i = X_{i\cdot}$ and $X_{ik} = \phi_{\xi_k}(x_i)$ for $1 \le i \le n, -D \cdot 2^D \le k \le D \cdot 2^D$. Then $[\mathbf{X}\mathbf{X}^\top]_{ij} = \mathbf{x}_i\mathbf{x}_j^\top = \sum_{k=-D\cdot 2^D}^{D\cdot 2^D} \phi_{\xi_k}(x_i)\phi_{\xi_k}(x_j)$, $[\mathbf{x}_0\mathbf{X}^\top]_j = \sum_{k=-D\cdot 2^D}^{D\cdot 2^D} \phi_{\xi_k}(x_0)\phi_{\xi_k}(x_j)$.

We want to show that $\hat{\mathbf{Y}}$ and $\hat{f}(\mathbf{x})$ are asymptotically equal as $D \to \infty$, which can be attained if we can show $2^{-D}\mathbf{x}_0\mathbf{X}^\top \to \boldsymbol{k}$ and $2^{-D}\mathbf{X}\mathbf{X}^\top \to \boldsymbol{K}$. By entry-wise comparison with our goal, it remains to show that for each given x_i and x_j, we have

$$2^{-D} \sum_{k=-D\cdot 2^D}^{D\cdot 2^D} \phi_{\xi_k}(x_i)\phi_{\xi_k}(x_j) \to \sqrt{\frac{\pi}{2}} e^{-\frac{(x_i-x_j)^2}{2}}$$

as $D \to \infty$.

Note that for each D, the set

$$\{\xi_k : k \in \mathbb{Z} \cap [-2^D D, 2^D D]\}$$

is equal to

$$\{a_{n,p} : n \in \mathbb{Z} \cap [-D, D-1], p \in \mathbb{Z} \cap [1, 2^D]\} \cup \{-D\},$$

where $a_{n,p} = n + 2^{-D}p$, as both of them characterized the fractions in $[-D, D]$ that are being used in the basis functions. Then one can facilitate the analysis by re-indexing the summation as follows:

$$\begin{aligned}
2^{-D} \sum_{k=-D\cdot 2^D}^{D\cdot 2^D} \phi_{\xi_k}(x_i)\phi_{\xi_k}(x_j) &= 2^{-D} \sum_{n=-D}^{D-1} \sum_{p=1}^{2^D} \phi_{a_{n,p}}(x_i)\phi_{a_{n,p}}(x_j) + 2^{-D}\phi_{-D}(x_i)\phi_{-D}(x_j) \\
&= 2^{-D} \sum_{n=-D}^{D-1} \sum_{p=1}^{2^D} e^{-(x_i-a_{n,p})^2} e^{-(x_j-a_{n,p})^2} + o(1) \\
&= e^{-\frac{1}{2}(x_i-x_j)^2} \sum_{n=-D}^{D-1} \sum_{p=1}^{2^D} 2^{-D} e^{-\frac{1}{2}(x_i+x_j-2a_{n,p})^2} + o(1) \\
&\overset{(*)}{=} e^{-\frac{1}{2}(x_i-x_j)^2} \sum_{n=-D}^{D-1} \left[\int_n^{n+1} e^{-\frac{1}{2}(x_i+x_j-2t)^2}\,\mathrm{d}t + O(2^{-D})\right] + o(1) \\
&= e^{-\frac{1}{2}(x_i-x_j)^2} \int_{-D}^{D} e^{-\frac{1}{2}(x_i+x_j-2t)^2}\,\mathrm{d}t + O(D2^{-D}) + o(1) \\
&= e^{-\frac{1}{2}(x_i-x_j)^2} \int_{-D}^{D} e^{-\frac{1}{2}(x_i+x_j-2t)^2}\,\mathrm{d}t + o(1),
\end{aligned}$$

where (*) follows from the error bound of a Riemann Sum, together with the fact that the integrand has a bounded slope (here, $O(g(D))$ means some function $f(D)$ that satisfies $\limsup_{D\to\infty} |f(D)|/g(D) < \infty$, and $o(g(D))$ means some function $f(D)$ that satisfies $\lim_{D\to\infty} f(D)/g(D) = 0$). Then as $D \to \infty$, we have

$$2^{-D} \sum_{k=-D\cdot 2^D}^{D\cdot 2^D} \phi_{\xi_k}(x_i)\phi_{\xi_k}(x_j) \to e^{-\frac{1}{2}(x_i-x_j)^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x_i+x_j-2t)^2}\,\mathrm{d}t = e^{-\frac{1}{2}(x_i-x_j)^2}\sqrt{\frac{\pi}{2}},$$

as desired. Here we used the property of the pdf of normal distributions, i.e. $\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{t-\mu}{1/2}\right)^2}\,\mathrm{d}t = \sqrt{2\pi(1/2)^2}$.