1. The ridge and lasso regression can be viewed as special cases of penalized least-squares

$$\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where p_λ is some penalization function. Consider an orthogonal design matrix, i.e., $\boldsymbol{X}^T\boldsymbol{X} = n\boldsymbol{I}_p$.

(a) Show that minimizing $Q(\boldsymbol{\beta})$ is equivalent to minimizing

$$\sum_{j=1}^{p}\left\{\frac{1}{2}(\widehat{\beta}_j - \beta_j)^2 + p_\lambda(|\beta_j|)\right\},$$

where $\widehat{\boldsymbol{\beta}}$ is the ordinary least-square estimate.

(b) Recall that in the case of orthogonal design matrix, lasso gives $\widehat{\beta}^{\text{lasso}} = \text{sign}(\widehat{\beta})(|\widehat{\beta}| - \lambda)_+$. As a result, the bais of lasso estimate is approximately λ for large true value β. To addess the bias, one idea is to use a penalty that tapers off as β becomes larger in absolute value, e.g., smoothly clipped absolute deviations (SCAD) penalty ($\gamma > 2$):

$$p_{\lambda,\gamma}(x) = \begin{cases} \lambda|x|, & \text{if } |x| \le \lambda, \\ \frac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma-1)}, & \text{if } \lambda < |x| \le \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2}, & \text{if } |x| > \gamma\lambda. \end{cases}$$

SCAD coincides with the lasso until $|x| = \lambda$, then it is quadratic, after which it remains constant. Use part (a), show that for the SCAD penalty, the solution under orthogonal design matrix is

$$\widehat{\beta}^{\text{SCAD}} = \begin{cases} \text{sign}(\widehat{\beta})(|\widehat{\beta}| - \lambda)_+, & \text{if } |\widehat{\beta}| \le 2\lambda, \\ \text{sign}(\widehat{\beta})\left[(\gamma-1)|\widehat{\beta}| - \gamma\lambda\right]/(\gamma-2), & \text{if } 2\lambda < |x| \le \gamma\lambda, \\ \widehat{\beta}, & \text{if } |x| > \gamma\lambda. \end{cases}$$

2. Let

$$\widehat{\beta}_\lambda = \underset{\beta}{\text{argmin}}\, g(\beta|\widehat{\beta},\lambda) = \underset{\beta}{\text{argmin}}\left\{\frac{1}{2}(\widehat{\beta} - \beta)^2 + p_\lambda(|\beta|)\right\},$$

for some penalty function $p_\lambda(\cdot)$. Following the convention, let $p'_\lambda(0) = p'_\lambda(0+)$ the right derivative at 0. Assume that $p_\lambda(\cdot)$ is nondescreasing and continuously differentiable on $[0,\infty)$, and the function $-\beta - p'_\lambda(\beta)$ is strictly unimodal on $[0,\infty)$.

(a) **[Sparsity]** Show that if $t_0 = \min_{\beta\ge 0}\{\beta + p'_\lambda(\beta)\} > 0$, then $\widehat{\beta}_\lambda = 0$ when $|\widehat{\beta}| \le t_0$.

(b) **[Unbiasedness]** Show that if $p'_\lambda(\beta) = 0$ for $|\beta| > t_1$, then $\widehat{\beta}_\lambda = \widehat{\beta}$ when $|\widehat{\beta}| \le t_1$ for large t_1.

3. Given data $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$ where $X_i \in \mathbb{R}^1$, consider the model

$$Y = f(X) + \epsilon,$$

where ϵ is a Gaussian noise with zero-mean and variance σ^2 and

$$f(x) = \sum_{i=-D\cdot 2^D}^{D\cdot 2^D} \theta_i \phi_{\xi_i}(x), \quad \text{with } \phi_{\xi_i}(x) = e^{-(x-\xi_i)^2}.$$

That is, we model Y as the sum of $2 \cdot D \cdot 2^D$ basis functions $\phi_{\xi_i}(\cdot)$, which are placed equally in the interval $[-D, D]$ with gaps 2^{-D}. Apparently the number of parameters is huge and we will definitely overfit the model. We learn the parameters using ridge regression with $\lambda = \sigma^2 \cdot 2^D$, that is, the cost function is given by

$$J(\boldsymbol{\theta}) = \|\boldsymbol{X\theta} - \boldsymbol{Y}\|^2 + \sigma^2 2^D \|\boldsymbol{\theta}\|^2.$$

Let $D \to \infty$, show that the prediction can be computed as

$$\widehat{f}(x) = \boldsymbol{k}(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{Y},$$

where $\boldsymbol{k} = (k_1, \ldots, k_N)$, $\boldsymbol{K} = \{K_{ij}\}_{0\le i,j\le N}$ with

$$k_i = k(x, X_i), \quad k_{i,j} = k(X_i, X_j)$$

and

$$k(x, y) = \sqrt{\frac{\pi}{2}} e^{-(x-y)^2/2}.$$