7. (**True or False**, 2 pts) The $L_2$ penalty in a ridge regression is equivalent to a Laplace prior on the weights.

**Solutions:** F

## 2. Ridge Regression [10 pts]

In class, we discussed $\ell_2$ penalized linear regression:

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2$$

where $X_i = [X_i^{(1)} \dots X_i^{(p)}]$.

a) Show that a closed form expression for the ridge estimator is $\widehat{\beta} = (\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^\top\boldsymbol{Y}$ where $\boldsymbol{A} = [X_1; \dots; X_n]$ and $\boldsymbol{Y} = [Y_1; ...; Y_n]$.

b) An advantage of ridge regression is that a unique solution always exists since $(\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I})$ is invertible. To be invertible, a matrix needs to be full rank. Argue that $(\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I})$ is full rank by characterizing its $p$ eigenvalues in terms of the singular values of $\boldsymbol{A}$ and $\lambda$.

a) (5 points) As in class, we start by writing the objective function in matrix form:

$$\begin{aligned}\mathcal{L} &= \|\boldsymbol{Y} - \boldsymbol{A\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\\ &= (\boldsymbol{Y} - \boldsymbol{A\beta})^\top(\boldsymbol{Y} - \boldsymbol{A\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}\end{aligned}$$

We will now compute the derivative of the objective function with respect to $\boldsymbol{\beta}$ using the following two vector differentiation rules: (1) $\partial\boldsymbol{b}^\top\boldsymbol{x}/\partial\boldsymbol{x} = \boldsymbol{b}$ and (2) $\partial\boldsymbol{x}^\top\boldsymbol{Bx}/\partial\boldsymbol{x} = (\boldsymbol{B} + \boldsymbol{B}^\top)\boldsymbol{x}$.

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} &= -2\boldsymbol{A}^\top(\boldsymbol{Y} - \boldsymbol{A\beta}) + 2\lambda\boldsymbol{\beta}\\ &= -2\boldsymbol{A}^\top\boldsymbol{Y} + 2\boldsymbol{A}^\top\boldsymbol{A\beta} + 2\lambda\boldsymbol{\beta}\end{aligned}$$

Since the derivative is zero at the minimizer $\hat{\boldsymbol{\beta}}$, we get

$$\begin{aligned}\boldsymbol{A}^\top\boldsymbol{Y} &= (\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I})\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^\top\boldsymbol{Y}\end{aligned}$$

b) (5 points) $\boldsymbol{A}^\top\boldsymbol{A}$ is a real symmetric matrix with real eigenvalues. It is also a positive semidefinite matrix, so all eigenvalues are nonnegative. (Equivalently, you can say that the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ are the square of the singular values of $\boldsymbol{A}$, and hence are nonnegative.) Let us denote these real nonnegative eigenvalues by $\{\nu_i\}_{i=1}^p$ (Notice that some of these can be zero).
Now notice that any eigenvector $\boldsymbol{v}_i$ of $\boldsymbol{A}^\top\boldsymbol{A}$ corresponding to eigenvalue $\nu_i$ is also an eigenvector of $\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I}$ with eigenvalue $\nu_i + \lambda$ since

$$(\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I})\boldsymbol{v}_i = \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}_i + \lambda\boldsymbol{v}_i = (\nu_i + \lambda)\boldsymbol{v}_i.$$

Since $\lambda > 0$, all eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I}$ are positive, so it is a full rank matrix and invertible.
**Note**: If you do not show that the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ are nonnegative, you only lose 2 points. If you do not argue why adding a positive constant $\lambda$ times identity matrix to $\boldsymbol{A}^\top\boldsymbol{A}$ makes the eigenvalues equal to $\nu_i + \lambda$, you lose 1 point.

### 2.2 Regularization: Ridge and Lasso Regression

**(a)**

We minimize $J_R(\beta)$ by setting its gradient w.r.t. $\beta$ to zero:

$$\begin{aligned}\nabla \mathrm{J}_R(\beta) = \nabla\left[(\mathbf{X}\beta - \mathbf{y})^\top(\mathbf{X}\beta-\mathbf{y}) + \lambda\beta^\top\beta\right] &= 0\\ 2\mathbf{X}^\top(\mathbf{X}\beta-\mathbf{y}) + 2\lambda\beta &= 0\\ \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}\right)\beta &= \mathbf{X}^\top\mathbf{y}\\ \beta^* &= \left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbb{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

**(b)**

Assuming $\mathbf{X}^\top\mathbf{X} = \mathbb{I}$, we minimize $\mathrm{J}_L(\beta)$ by setting its gradient w.r.t. $\beta$ to zero:

$$\begin{aligned}\nabla \mathrm{J}_R(\beta) = \nabla\left[(\mathbf{X}\beta - \mathbf{y})^\top(\mathbf{X}\beta-\mathbf{y}) + \lambda\|\beta\|_1\right] &= 0\\ 2\mathbf{X}^\top(\mathbf{X}\beta-\mathbf{y}) + \lambda\nabla\|\beta\|_1 &= 0\\ 2\beta + \lambda\nabla\|\beta\|_1 &= 2\mathbf{X}^\top\mathbf{y}\end{aligned}$$

Let us consider the derivative w.r.t. $\beta_a$ for some $1 \leq a \leq M$:

$$2\beta_a + \lambda\frac{d}{d\beta_a}|\beta_a| \;=\; 2\sum_i x_{i,a}y_i$$

There are 3 cases to consider: either the optimal value of $\beta_a$ is $> 0$, $< 0$, or $= 0$. If we assume the optimal $\beta_a > 0$, then $\frac{d}{d\beta_a}|\beta_a| = 1$ and we have

$$\begin{aligned}2\beta_a + \lambda &= 2\sum_i x_{i,a}y_i\\ \beta_a &= \left(\sum_i x_{i,a}y_i\right) - \frac{\lambda}{2}\end{aligned}$$

provided that $(\sum_i x_{i,a}y_i) > \frac{\lambda}{2}$. If we assume the optimal $\beta_a < 0$, then $\frac{d}{d\beta_a}|\beta_a| = -1$ and we have

$$\begin{aligned}2\beta_a - \lambda &= 2\sum_i x_{i,a}y_i\\ \beta_a &= \left(\sum_i x_{i,a}y_i\right) + \frac{\lambda}{2}\end{aligned}$$

provided that $(\sum_i x_{i,a}y_i) < -\frac{\lambda}{2}$. For the final case where the optimal $\beta_a = 0$, there is no closed form solution to the gradient $\frac{d}{d\beta_a}|\beta_a|$. We observe this corresponds to situations where $-\frac{\lambda}{2} \leq (\sum_i x_{i,a}y_i) \leq \frac{\lambda}{2}$. In other words, $-\frac{\lambda}{2} \leq (\sum_i x_{i,a}y_i) \leq \frac{\lambda}{2}$ implies $\beta_a = 0$.

Hence the optimal $\beta^*$ is

$$\beta_a^* \;=\; \begin{cases}(\sum_i x_{i,a}y_i) - \frac{\lambda}{2} & \text{if } (\sum_i x_{i,a}y_i) > \frac{\lambda}{2}\\ (\sum_i x_{i,a}y_i) + \frac{\lambda}{2} & \text{if } (\sum_i x_{i,a}y_i) < -\frac{\lambda}{2}\\ 0 & \text{otherwise.}\end{cases}$$

**(c)**

Assuming $\mathbf{X}^\top \mathbf{X} = \mathbb{I}$, vanilla linear regression implies $\beta_a^* = \sum_i x_{i,a} y_i$, ridge regression implies $\beta_a^* = \frac{1}{\lambda} \sum_i x_{i,a} y_i$, and Lasso regression implies

$$\beta_a^* = \begin{cases} \left(\sum_i x_{i,a} y_i\right) - \frac{\lambda}{2} & \text{if } \left(\sum_i x_{i,a} y_i\right) > \frac{\lambda}{2} \\ \left(\sum_i x_{i,a} y_i\right) + \frac{\lambda}{2} & \text{if } \left(\sum_i x_{i,a} y_i\right) < -\frac{\lambda}{2} \\ 0 & \text{otherwise.} \end{cases}$$

With respect to vanilla linear regression, ridge regression shrinks the $\beta^*$ estimates by a factor of $\frac{1}{\lambda}$, whereas Lasso regression translates the $\beta^*$ estimates by a distance $\frac{\lambda}{2}$ towards zero.

## 1.1 Ridge regression

Starting from our true model $\mathbf{y} = \mathbf{X}\theta + \epsilon$, we express $\hat{\theta}$ in terms of $\epsilon$ and $\theta$:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\theta + \epsilon \\ \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{y} &= \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \left(\mathbf{X}\theta + \epsilon\right) \\ \hat{\theta} &= \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{X}\theta + \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \epsilon. \end{aligned}$$

Because $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$, it follows from the hint that

$$\begin{aligned} \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \epsilon &\sim \mathcal{N}\left(0, \sigma^2 \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \left(\left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top\right)^\top\right) \\ &= \mathcal{N}\left(0, \sigma^2 \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1}\right). \end{aligned}$$

Hence

$$\hat{\theta} \sim \mathcal{N}\left(\left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{X}\theta,\ \sigma^2 \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1}\right),$$

in other words $\hat{\theta}$ has a Gaussian distribution with mean and covariance

$$\begin{aligned} E\left[\hat{\theta}\right] = \mu &= \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{X}\theta \neq \theta \\ \Sigma &= \sigma^2 \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1}, \end{aligned}$$

implying that ridge regression is biased.

## 1.2 Extra features

Let $\beta_{new}$ be the parameter corresponding to the $n \times 1$ vector of new features $X_{new}$, and define

$$B = \begin{bmatrix} \beta \\ \beta_{new} \end{bmatrix}$$

to be the original vector $\beta$ concatenated with $\beta_{new}$. Also, let $\mathrm{J}_{new}(B) = \left(\mathbf{X}_{new} B - y\right)^\top \left(\mathbf{X}_{new} B - y\right)$ be the squared error for the augmented feature values $\mathbf{X}_{new}$.

Let's derive a few facts beforehand. The assumption $\mathbf{X}_{new}^\top \mathbf{X}_{new} = I$ implies that the columns of $\mathbf{X}_{new}$ are orthonormal, which in turn implies:

1. $\mathbf{X}^\top \mathbf{X} = I$

2. $\mathbf{X}^\top X_{new} = \vec{0}$ where $\vec{0}$ is the zero vector

3. $X_{new}^\top X = 1$

We know that the minimizer for $\mathrm{J}_{new}(B)$ is

$$\begin{aligned}\hat{B} &= \left(\mathbf{X}_{new}^\top \mathbf{X}_{new}\right)^{-1} \mathbf{X}_{new}^\top \mathbf{y} \\ &= \mathbf{X}_{new}^\top \mathbf{y} \quad (\text{because } \mathbf{X}_{new}^\top \mathbf{X}_{new} = I),\end{aligned}$$

while the minimizer for the original problem $\mathrm{J}_1(\beta)$ is

$$\begin{aligned}\hat{\beta} &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \mathbf{X}^\top \mathbf{y} \quad (\text{because } \mathbf{X}^\top \mathbf{X} = I).\end{aligned}$$

Given the above facts, the minimum value of $\mathrm{J}_{new}(B)$ is given by

$$\begin{aligned}\mathrm{J}_{new}\left(\hat{B}\right) &= \left(\mathbf{X}_{new}\hat{B} - \mathbf{y}\right)^\top \left(\mathbf{X}_{new}\hat{B} - \mathbf{y}\right) \\ &= \left(\mathbf{X}_{new}\mathbf{X}_{new}^\top \mathbf{y} - \mathbf{y}\right)^\top \left(\mathbf{X}_{new}\mathbf{X}_{new}^\top \mathbf{y} - \mathbf{y}\right) \\ &= \left(\left(\mathbf{X}\mathbf{X}^\top + X_{new}X_{new}^\top\right)\mathbf{y} - \mathbf{y}\right)^\top \left(\left(\mathbf{X}\mathbf{X}^\top + X_{new}X_{new}^\top\right)\mathbf{y} - \mathbf{y}\right) \\ &= \left(\left(\mathbf{X}\mathbf{X}^\top \mathbf{y} - \mathbf{y}\right) + X_{new}X_{new}^\top \mathbf{y}\right)^\top \left(\left(\mathbf{X}\mathbf{X}^\top \mathbf{y} - \mathbf{y}\right) + X_{new}X_{new}^\top \mathbf{y}\right) \\ &= \left(\mathbf{X}\mathbf{X}^\top \mathbf{y} - \mathbf{y}\right)^\top \left(\mathbf{X}\mathbf{X}^\top \mathbf{y} - \mathbf{y}\right) + 2\left(\mathbf{X}\mathbf{X}^\top \mathbf{y} - \mathbf{y}\right)^\top \left(X_{new}X_{new}^\top \mathbf{y}\right) + \left(X_{new}X_{new}^\top \mathbf{y}\right)^\top \left(X_{new}X_{new}^\top \mathbf{y}\right).\end{aligned}$$

Observing that $\mathrm{J}_1\left(\hat{\beta}\right) = \left(\mathbf{X}\mathbf{X}^\top \mathbf{y} - \mathbf{y}\right)^\top \left(\mathbf{X}\mathbf{X}^\top \mathbf{y} - \mathbf{y}\right)$, we get

$$\begin{aligned}\mathrm{J}_{new}\left(\hat{B}\right) &= \mathrm{J}_1\left(\hat{\beta}\right) + 2\left(\mathbf{X}\mathbf{X}^\top \mathbf{y} - \mathbf{y}\right)^\top \left(X_{new}X_{new}^\top \mathbf{y}\right) + \left(X_{new}X_{new}^\top \mathbf{y}\right)^\top \left(X_{new}X_{new}^\top \mathbf{y}\right) \\ &= \mathrm{J}_1\left(\hat{\beta}\right) + 2\left(\mathbf{X}\mathbf{X}^\top \mathbf{y}\right)^\top \left(X_{new}X_{new}^\top \mathbf{y}\right) - 2\mathbf{y}^\top \left(X_{new}X_{new}^\top \mathbf{y}\right) + \left(X_{new}X_{new}^\top \mathbf{y}\right)^\top \left(X_{new}X_{new}^\top \mathbf{y}\right) \\ &= \mathrm{J}_1\left(\hat{\beta}\right) + 2\left(\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top X_{new}X_{new}^\top \mathbf{y}\right) - 2\left(\mathbf{y}^\top X_{new}X_{new}^\top \mathbf{y}\right) + \left(\mathbf{y}^\top X_{new}X_{new}^\top X_{new}X_{new}^\top \mathbf{y}\right) \\ &= \mathrm{J}_1\left(\hat{\beta}\right) + 2\left(\mathbf{y}^\top \mathbf{X}\vec{0}X_{new}^\top \mathbf{y}\right) - 2\left(\mathbf{y}^\top X_{new}X_{new}^\top \mathbf{y}\right) + \left(\mathbf{y}^\top X_{new}X_{new}^\top \mathbf{y}\right) \\ &= \mathrm{J}_1\left(\hat{\beta}\right) - \mathbf{y}^\top X_{new}X_{new}^\top \mathbf{y} \\ &= \mathrm{J}_1\left(\hat{\beta}\right) - \left(X_{new}^\top \mathbf{y}\right)^\top \left(X_{new}^\top \mathbf{y}\right) \\ &\leq \mathrm{J}_1\left(\hat{\beta}\right).\end{aligned}$$

In other words, our new feature $X_{new}$ allows us to decrease (or at least maintain) the minimized objective value.

## Q4) Kernel Regression (Pengtao)

### Kernel Regression (10 points)

Consider local linear regression where the predicted output value of $x$ is $\hat{f}(x) = \hat{\alpha} + \hat{\beta}x$, where

$$\hat{\alpha}, \hat{\beta} = \mathrm{argmin}_{\alpha,\beta} \sum_{i=1}^{n} w_i(x)(y_i - \alpha - \beta x_i)^2 \tag{17}$$

where $w_i(x) = K(\frac{x-x_i}{h}) / \sum_{i=1}^{n} K(\frac{x-x_i}{h})$.

1. Show that the objective function can be re-written as

$$(\boldsymbol{y} - \boldsymbol{Ba})^\top \Omega(x)(\boldsymbol{y} - \boldsymbol{Ba})$$

where $\boldsymbol{y} = [y_1\ y_2\ \dots y_n]^\top$, $\boldsymbol{a} = [\alpha\ \beta]^\top$, $\boldsymbol{B} = [1\ x_1;\ 1\ x_2;\ \dots;\ 1\ x_n]$ and $\Omega(x)$ is a diagonal matrix with diagonal $[w_1(x)\ w_2(x)\ \dots\ w_n(x)]$ .

**(4 pts)**

$$\begin{aligned}
&\sum_{i=1}^{n} w_i(x)(y_i - \alpha - \beta x_i)^2 \\
&= \sum_{i=1}^{n} w_i(x)(y_i - [1\ x_i]\mathbf{a})^\top (y_i - [1\ x_i]\mathbf{a}) \quad (\text{where } \mathbf{a} = [\alpha\ \beta]^\top) \\
&= \sum_{i=1}^{n} w_i(x)(\mathbf{y}_i - \mathbf{B}_i\mathbf{a})^\top (\mathbf{y}_i - \mathbf{B}_i\mathbf{a}) \quad (\text{where } \mathbf{y} = [y_1\ y_2\ \dots y_n]^\top,\ \mathbf{B} = [1\ x_1;\ 1\ x_2;\ \dots;\ 1\ x_n], \\
&\qquad\qquad \mathbf{y}_i \text{ is the } i\text{th row of } \mathbf{y}, \mathbf{B}_i \text{ is the } i\text{th row of } \mathbf{B}) \\
&= (\mathbf{y} - \mathbf{B}\mathbf{a})^\top \Omega(x)(\mathbf{y} - \mathbf{B}\mathbf{a}) \quad (\text{where } \Omega(x) \text{ is a diagonal matrix with diagonal } [w_1(x)\ w_2(x)\ \dots\ w_n(x)])
\end{aligned} \tag{18}$$

2. Show that $\hat{f}(x)$ is a linear combination of $\{y_i\}_{i=1}^n$, namely $\hat{f}(x)$ can be written as $\hat{f}(x) = \sum_{i=1}^{n} \ell_i(x) y_i = \boldsymbol{\ell}(x)^\top \boldsymbol{y}$, where $\ell_i(x)$ is some quantity defined over $x$ and $\boldsymbol{\ell}(x) = [\ell_1(x)\ \ell_2(x)\ \dots\ \ell_n(x)]^\top$.

**(6 pts)**

First, we solve $\hat{\mathbf{a}} = \text{argmin}_{\mathbf{a}}\ \ (\mathbf{y} - \mathbf{B}\mathbf{a})^\top \Omega(x)(\mathbf{y} - \mathbf{B}\mathbf{a})$. Taking the derivative of $(\mathbf{y} - \mathbf{B}\mathbf{a})^\top \Omega(x)(\mathbf{y} - \mathbf{B}\mathbf{a})$ w.r.t $\mathbf{a}$ and setting the derivative to zero, we get

$$-\mathbf{B}^\top \Omega(x)(\mathbf{y} - \mathbf{B}\mathbf{a}) = 0 \tag{19}$$

Solving for $\mathbf{a}$, we have

$$\hat{\mathbf{a}} = (\mathbf{B}^\top \Omega(x)\mathbf{B})^{-1}\mathbf{B}^\top \Omega(x)\mathbf{y} \tag{20}$$

$$\begin{aligned}
\hat{f}(x) &= b^\top(x)\hat{\mathbf{a}} \\
&= b^\top(x)(\mathbf{B}^\top \Omega(x)\mathbf{B})^{-1}\mathbf{B}^\top \Omega(x)\mathbf{y} \\
&= \boldsymbol{\ell}(x)^\top \mathbf{y}
\end{aligned} \tag{21}$$

where $b^\top(x) = [1 \quad x]$ and $\boldsymbol{\ell}(x)^\top = b^\top(x)(\mathbf{B}^\top \Omega(x)\mathbf{B})^{-1}\mathbf{B}^\top \Omega(x)$

## Kernelized Ridge Regression (10 points)

The nonparametric kernel regression does a local fit around the test point. Lets now investigate the use of kernels for regression in another way. Similar to the kernel trick for SVMs, we can apply the kernel trick in regression as follows. (However, note that this produces a global fit to the data.)

Consider the ridge regression problem where we have a set of data points $\{\mathbf{x}_i\}_{i=1}^N$ and corresponding response values $\{y_i\}_{i=1}^N$. To achieve better performance, we use a feature mapping function $\phi$ to map the original $d$-dimensional feature vector $\mathbf{x}$ to a new $D$-dimensional feature vector $\hat{\mathbf{x}} = \phi(\mathbf{x})$, where $D \gg d$. Let $\mathbf{\Phi} \in \mathbb{R}^{N \times D}$ denote the design matrix, whose $i$th row contains the new feature vector $\hat{\mathbf{x}}_i^\top$ of the $i$th data point. Let $\mathbf{y}$ denote the response value vector whose $i$th component is the response value $y_i$ of the $i$th point.

For ridge regression, the objective function is $J(\boldsymbol{\beta}) = \frac{1}{2}||\mathbf{y} - \mathbf{\Phi}\boldsymbol{\beta}||^2 + \frac{\lambda}{2}||\boldsymbol{\beta}||^2$, where $\lambda$ is the tradeoff parameter. Let $\boldsymbol{\beta}^*$ denote the solution to the ridge regression. In the following steps, we are going to derive the kernel ridge regression.

(a) First, show that $\boldsymbol{\beta}^*$ is in the space spanned by rows in $\mathbf{\Phi}$, i.e., $\boldsymbol{\beta}^*$ can be written in the form $\boldsymbol{\beta}^* = \mathbf{\Phi}^\top \boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^* \in \mathbb{R}^{N \times 1}$. (Hint: $\boldsymbol{\beta}$ can be decomposed into $\boldsymbol{\beta} = \boldsymbol{\beta}_{||} + \boldsymbol{\beta}_{\perp}$, where $\boldsymbol{\beta}_{\perp}$ is orthogonal to the span of rows in $\mathbf{\Phi}$ and $\boldsymbol{\beta}_{||}$ lies in the span of rows in $\mathbf{\Phi}$.)

**(4 pts)**

$$\begin{aligned} J(\boldsymbol{\beta}) &= J(\boldsymbol{\beta}_{||} + \boldsymbol{\beta}_{\perp}) \\ &= \tfrac{1}{2}||\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\beta}_{||} - \boldsymbol{\Phi}\boldsymbol{\beta}_{\perp}||^2 + \lambda||\boldsymbol{\beta}_{\perp}||^2 + \lambda||\boldsymbol{\beta}_{||}||^2 \\ &= \tfrac{1}{2}||\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\beta}_{||}||^2 + \lambda||\boldsymbol{\beta}_{\perp}||^2 + \lambda||\boldsymbol{\beta}_{||}||^2 \end{aligned} \tag{22}$$

To minimize this objective, $\boldsymbol{\beta}_{\perp}$ needs to be $\mathbf{0}$. Thereby, $\boldsymbol{\beta}^*$ lies in the span of rows in $\boldsymbol{\Phi}$.

(b) In (a), we have proved that $\boldsymbol{\beta}^* = \boldsymbol{\Phi}^{\top}\boldsymbol{\alpha}^*$. In this step, show that $\boldsymbol{\alpha}^* = (\boldsymbol{\Phi}\boldsymbol{\Phi}^{\top} + \lambda\mathbf{I})^{-1}\mathbf{y}$.

**(4 pts)**

Plug in $\boldsymbol{\beta} = \boldsymbol{\Phi}^{\top}\boldsymbol{\alpha}$ into $J(\boldsymbol{\beta})$

$$\begin{aligned} &\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \\ &= \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha}\|_2^2 + \lambda\boldsymbol{\alpha}^T\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha} \end{aligned} \tag{23}$$

Taking derive of Eq.(23) w.r.t $\boldsymbol{\alpha}$ and setting it to zero, we get

$$-2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha}) + 2\lambda\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \tag{24}$$

Solving for $\boldsymbol{\alpha}$, we get

$$\boldsymbol{\alpha} = (\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda I)^{-1}\mathbf{y} \tag{25}$$

(c) In practice, it is very hard to design the feature mapping function $\phi$. Even if we can design it, computing the inner product $\phi(\mathbf{x})^{\top}\phi(\mathbf{x}')$ between two points can be costly. Instead, we can use a kernel function $k(\mathbf{x}, \mathbf{x}')$ to implicitly compute the inner product of high dimensional features, i.e., $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\top}\phi(\mathbf{x}')$. Basically, kernelization means using the kernel function to replace inner products. In this step, given the kernel function $k(\mathbf{x}, \mathbf{x}')$, where $x$ and $x'$ are the original feature vectors of dimension $d$, try to kernelize the ridge regression. You need to consider both training and testing. (Hint: In training phase, you need to replace all inner produces in $\boldsymbol{\alpha}^*$ with kernel function. In testing phase, giving a new test point $\mathbf{x}$, you need to compute $\phi(\mathbf{x})^{\top}\boldsymbol{\beta}^*$. Replace all inner products in $\phi(\mathbf{x})^{\top}\boldsymbol{\beta}^*$ with kernel function.)

**(2 pts)**

In the training phase, we need to compute $\boldsymbol{\alpha}$. Let $\mathbf{K}$ denote a matrix where $K_{ij} = k(\mathbf{x}_{\mathrm{i}}, \mathbf{x}_{\mathrm{j}})$. It can be easily checked that $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}$. So, $\boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$. In the testing phase, we need to compute $\phi(\mathbf{x})^{\top}\boldsymbol{\beta}^* = \phi(\mathbf{x})^{\top}\boldsymbol{\Phi}^{\top}\boldsymbol{\alpha} = \sum_{i=1}^{N}\alpha_i k(x_i, x)$