I have neither given nor received unauthorized assistance on this exam.

________________________________________________________
Name Signed                        Date
________________________________________________________
Name Printed

This is an extremely long exam. Do as much of it as you can in 2 hours. I'll score it 10 points per page. Some pages will go (much) faster than others, and you'll be wise to do them first.
1. Consider the $p = 3$ linear prediction problem with $N = 5$ and training data

\[
\begin{bmatrix}
1 & 0 & 1 \\
0 & 1 & 1 \\
0 & -1 & 1 \\
-1 & 0 & 1 \\
0 & 0 & -4
\end{bmatrix}
\text{ and } \begin{bmatrix}
2 \\
3 \\
-1 \\
-1 \\
-3
\end{bmatrix}
\]

In answering the following, you may use the notation that the $j$th column of $X$ is $x_j$. 

a) Find the fitted OLS coefficient vector $\hat{\beta}_{\text{OLS}}$.

b) For $\lambda = 10$ find the fitted coefficient vector minimizing

\[
\left( Y - X \text{ diag}(c) \hat{\beta}_{\text{OLS}} \right)' \left( Y - X \text{ diag}(c) \hat{\beta}_{\text{OLS}} \right) + \lambda I'c
\]

over choices of $c \in \mathbb{R}^3$ with non-negative entries.
c) For \( \lambda > 0 \) find the fitted ridge coefficient vector, \( \beta_{\lambda}^{\text{ridge}}. \)

d) For \( \lambda > 0 \) find a fitted coefficient vector \( \hat{\beta}_{\lambda}^{*} \) minimizing \( (Y - Xb)'(Y - Xb) + \lambda \left( b_2^2 + b_3^2 \right) \) as a function of \( b \in \mathbb{R}^3. \)
e) Carefully specify the entire Least Angle Regression path of either $\hat{Y}$ or $\hat{\beta}$ values.
2. As it turns out

\[
\begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\
0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{20}} \\
0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{20}} \\
-\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\
0 & 0 & -\frac{4}{\sqrt{20}}
\end{pmatrix}
\begin{pmatrix}
\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{20}} \\
0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\
0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\
-\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{20}} \\
0 & 0 & -\frac{4}{\sqrt{20}}
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{20}} \\
0 & 0 & \frac{1}{\sqrt{2}} \\
0 & \frac{1}{2} & 0 \\
0 & 0 & \frac{1}{\sqrt{2}}
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

Consider a \( p = 3 \) linear prediction problem where the matrix of training inputs, \( \mathbf{X} \), is the matrix on the left above and \( \mathbf{Y}' = (4, 2, 2, 0, 2) \).

a) Find the single principal component (\( M = 1 \)) fitted coefficient vector \( \hat{\beta}^{\text{PCR}} \).
b) Find the single component ($M = 1$) partial least squares vector of predictions, $\hat{Y}_{\text{PLS}}$. (You must provide a numerical answer.)
3. Suppose that $C \subseteq \Re^p$. This question is about "kernels" ($K(x, z)$), non-negative definite functions $C \times C \rightarrow \Re$ and corresponding RKHS's.

a) Show that for $\phi : C \rightarrow \Re$, the function $K(x, z) = \phi(x)\phi(z)$ is a valid kernel. (You must show that for distinct $x_1, x_2, \ldots, x_M$, the $M \times M$ matrix $K = (K(x_i, x_j))$ is non-negative definite.

b) Show that for two kernels $K_1(x, z)$ and $K_2(x, z)$ and two positive constants $c_1$ and $c_2$, the function $c_1K_1(x, z) + c_2K_2(x, z)$ is a kernel.

c) By virtue of a) and b), the functions $K_1(x, z) = 1 + xz$ and $K_2(x, z) = 1 + 2xz$ are both kernels on $[-1, 1]^2$. They produce different RKHS's. Show these are different.
4. Consider the RKHS of functions on $[-2, 2]$ defined by the kernel $K(x, z) = 1 + xz \cdot \exp(x + z)$ on $[-2, 2]^2$. You may take as given the fact that all functions $K(x, c)$ of $x$ for $c \in [-2, 2]$ belong to this RKHS, $\mathcal{H}_K$.

a) Show that the functions $g(x) = 1$ and $h(x) = x \exp(x)$ both belong to $\mathcal{H}_K$.

b) Determine whether or not $g$ and $h$ are orthonormal. If they are not, find an orthonormal basis for the span of $\{g, h\}$ in $\mathcal{H}_K$. 

5. Suppose that $P$ is such that $x$ has pdf $f(x) = \frac{3}{2} I\left[0 < x < \frac{1}{2}\right] + \frac{1}{2} I\left[\frac{1}{2} < x < 1\right]$ on $[0,1]$ and the conditional distribution of $y|x$ is $N(x,1)$. Suppose training data $(x_i, y_i)$ for $i = 1, 2, \ldots, N$ are iid $P$ and that with $\phi$ the standard normal pdf, one uses the Nadaraya-Watson estimator for $E[y|x = .5] = .5$,

$$\hat{f}(.5) = \frac{\sum_{i=1}^{N} y_i \phi(.5 - x_i)}{\sum_{i=1}^{N} \phi(.5 - x_i)}$$

Use the law of large numbers and the continuity of the ratio function and write out the (in probability) limit for $\hat{f}(.5)$ in terms of a ratio of two definite integrals and argue that the limit is not .5.
6. Consider a toy problem where one is to employ locally weighted straight line regression smoothing based on the Epanechnikov quadratic kernel in a \( p = 1 \) context with training data below. Using a bandwidth of \( \lambda = .5 \), give a small (augmented) data set for which ordinary simple linear regression (unweighted least squares) will produce the smoothed prediction at \( x = 0 \) (that is, \( \hat{f}_3(0) \)) for the original training data.

<table>
<thead>
<tr>
<th>( x )</th>
<th>-1.0</th>
<th>-0.75</th>
<th>-0.50</th>
<th>-0.25</th>
<th>0</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
7. Consider again the toy data set in Problem 6. Set up an $X$ matrix for an ordinary multiple linear regression that could be used to fit a linear regression spline with knots at $\xi_1 = -.5, \xi_2 = 0$, and $\xi_3 = .5$. For your set-up, what linear combination of fitted regression parameters produces the prediction at $x = 0$?
8. Consider a $p = 2$ prediction problem with continuous univariate output $y$. Two possible methods of prediction are under consideration, namely

1. a neural net with single hidden layer and $M = 2$ hidden nodes (and single output node) using $\sigma(u) = 1/(1 + \exp(u))$ and $g(v) = v$, and

2. a projection pursuit regression predictor with $M = 2$ summands $g_m(w_m'x)$ (based on cubic smoothing splines).

a) Argue carefully that in general, possibility 2 provides more flexibility in fitting than possibility 1.

b) Note that unit vectors in $\mathbb{R}^2$ can be parameterized by a single real variable $\theta \in (-\pi, \pi]$. How would you go about choosing a version of possibility 2 that might be expected to provide only "about as much flexibility in fitting" as possibility 1? (This will have to amount to some speculation on your part, but make a sensible suggestion based on "parameter counts.")