Midterm

Justify all your answers. Full credit will not be given to any answer that is not adequately justified. You can use results from the notes and other material as long as you explain why they apply.

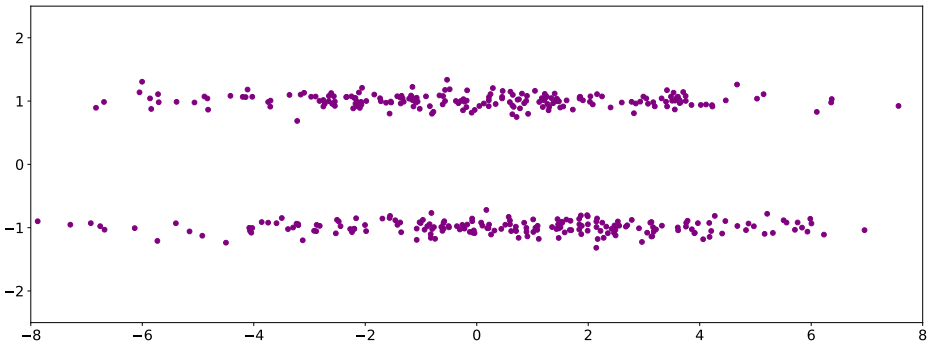
You are not allowed to communicate with anyone about the midterm while you are doing it. We have preferred to prioritize making your life as easy as possible from the logistical point of view, instead of making the timing very tight to avoid cheating. Please don't abuse this. It is not fair to your fellow students.

Submit the midterm through Gradescope, just like the homework. Please remember to start each problem on a separate page (just like the homework).

1. *Short questions (*20 points*).*

   a. We are interested in clustering these 2D data. Is PCA the right way of performing dimensionality reduction to 1D for this task? (5 points)

   Doing PCA we will select the horizontal axis as the first principal direction since most of the variance of the data is explained by projecting along this axis but then we will not be able to separate the initial two clusters which were separated clearly in 2D. So PCA is not the right way of performing dimensionality reduction to 1D for this task.



   b. We are interested in computing the singular values of a matrix $A \in \mathbb{R}^{p \times p}$, but we cannot observe it directly. We can only observe its output for some inputs of our choice. A popular technique in such settings is to choose random inputs. Explain how to compute the singular values of A from the random vector $\tilde{y} := A\tilde{x}$, where $\tilde{x}$ is a zero mean random vector with covariance matrix equal to the identity. (5 points)

   Using theorem 8.6 from PCA notes, we have $\Sigma_{\tilde{y}} = A\Sigma_{\tilde{x}}A^T = AA^T$, then with random inputs we can build the covariance matrix using its outputs, which is a square matrix. We can then perform an eigendecomposition which gives us the eigenvalues of AA^T, which is positive definite, the singular value of A which are non negative will be the square roots of these eigenvalues.

   c. Explain how performing PCA on the feature matrix of a linear regression problem can help you decide whether to apply OLS or ridge regression. (5 points)

   Performing PCA on the feature matrix gives us the eigenvalues of the covariance matrix and if some eigenvalues are very small compared to the others, it is the sign of correlations between features. Then the covariance matrix might be not full rank and the OLS estimator coefficients might blow

up, also the small singular values might cause high variance of the OLS coefficients if there is not a lot of samples. We might want then decide to use ridge regression instead to neutralize the high variances due to the small singular values.

d. For a fixed feature matrix and a fixed response vector, can the training error of OLS be larger than the training error of ridge regression for some value of the regularization parameter $\lambda > 0$? (5 points)

if we compare OLS and ridge regression training errors:

$$\text{OLS} = \min \| \left[y - X^T \beta \right] \|_2^2$$

$$\text{RR} = \min \| \begin{bmatrix} y - X^T \beta \\ \sqrt{\lambda} I \end{bmatrix} \|_2^2$$

$$\| \left[y - X^T \beta \right] \|_2^2 \leq \| \begin{bmatrix} y - X^T \beta \\ \sqrt{\lambda} I \end{bmatrix} \|_2^2 = \| \left[y - X^T \beta \right] \|_2^2 + \lambda \| I \|_2^2$$

Taking the minimum on both sides shows that for any $\lambda > 0$, we have that the training error of OLS cannot be larger than the training error of ridge regression.

2. *Normalization (*30 points*).* In this problem we study the effect of normalizing by the standard deviation before performing principal component analysis. Let $\tilde{x}$ be a zero-mean 3-dimensional random vector with covariance matrix

$$\Sigma_{\tilde{x}} := \begin{bmatrix} 100 & 25 & 0 \\ 25 & 400 & 0 \\ 0 & 0 & 0.16 \end{bmatrix}. \tag{1}$$

We define the normalized vector $\tilde{y}$ as

$$\tilde{y}[i] := \frac{\tilde{x}[i]}{\sqrt{\text{Var}(\tilde{x}[i])}} \quad 1 \leq i \leq 3. \tag{2}$$

a. Compute the covariance matrix of $\tilde{y}$. (10 points)

$$\begin{aligned} \text{Var}(\tilde{y}[i]) &= \text{Var}(\frac{\tilde{x}[i]}{\sqrt{\text{Var}(\tilde{x}[i])}}) \quad 1 \leq i \leq 3 \\ &= \frac{1}{\text{Var}(\tilde{x}[i])}\text{Var}(\tilde{x}[i]) \\ &= 1 \\ \text{Cov}(\tilde{y}[i]\tilde{y}[j]) &= \text{Cov}(\frac{\tilde{x}[i]}{\sqrt{\text{Var}(\tilde{x}[i])}}\frac{\tilde{x}[j]}{\sqrt{\text{Var}(\tilde{x}[j])}}) \; i \neq j \quad 1 \leq i, j \leq 3 \\ &= \frac{1}{\sqrt{\text{Var}(\tilde{x}[i])\text{Var}(\tilde{x}[j])}}\text{Cov}(\tilde{x}[i]\tilde{x}[j]) \end{aligned}$$

Using the covariance of $\tilde{x}$ we obtain the covariance of $\tilde{y}$:

$$\Sigma_{\tilde{y}} := \begin{bmatrix} 1 & 0.125 & 0 \\ 0.125 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

b. Is the directional variance of $\tilde{y}$ equal to one in every direction? (5 points) The directional variance will be the same in every direction if we have for covariance of $\tilde{y}$ the identity matrix which is not the case, so it is not the same in every direction.

c. We decide to reduce the dimensionality of $\tilde{x}$ and $\tilde{y}$ to two dimensions using PCA. Report what directions are selected for each of the random vectors. (Feel free to use a computer for your calculations, but explain what you are doing.) (10 points) We perform the eigendecomposition of the covariance of $\tilde{x}$ and found its eigenvalues and eigenvectors:

$$\begin{aligned} \Sigma_{\tilde{x}} &= U_x D_x U_x^T \\ D_x &= \begin{bmatrix} 0.16 & 0 & 0 \\ 0 & 97.93 & 0 \\ 0 & 0 & 402.06 \end{bmatrix} \\ &= [\lambda_3 \; \lambda_2 \; \lambda_1] \\ U_x &= \begin{bmatrix} 0 & -0.996 & 0.0824 \\ 0 & 0.0824 & 0.996 \\ 1 & 0 & 0 \end{bmatrix} \\ &= [u_3 \; u_2 \; u_1] \end{aligned}$$

Using theorem 6.1 from the PCA notes the first eigenvector u_1 is the direction of highest variance, which is equal to corresponding eigenvalue λ_1. In directions orthogonal to u_1 the maximum variance is attained by the second eigenvector u_2, and equals the corresponding eigenvalue λ_2. And the last eigenvalue 0.16 is very small compared to the two others so most of the variance is captured by the two eigenvectors u_1 and u_2. If we project the data using the two eigenvectors u_1 and u_2 we will reduce the dimensions from three to two while conserving most of the variance. Also as a result of theorem7.1 from PCA notes, projection along these directions achieves the highest variances compared to any other directions.

Similarly for $\tilde{y}$ we compute the eigendecomposition of its covariance matrix and find:

$$\begin{aligned}
\Sigma_{\tilde{y}} &= U_y D_y U_y^T \\
D_y &= \begin{bmatrix} 0.875 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1.125 \end{bmatrix} \\
&= [\mu_3 \ \mu_2 \ \mu_1] \\
U_y &= \begin{bmatrix} -0.7071 & 0 & 0.7071 \\ 0.7071 & 0 & 0.7071 \\ 0 & 1 & 0 \end{bmatrix} \\
&= [v_3 \ v_2 \ v_1]
\end{aligned}$$

Using theorem 6.1 from PCA notes, we select for directions v_1 and v_2 corresponding to the largest two eigenvalues and project the data along these two directions.

d. Explain which of the two options for dimensionality reduction to 2D would make more sense in each of the following situations and why that is the case: (1) The entries of $\tilde{x}$ represent the weight (in kilograms), heart rate (in beats per minute), and height (in meters) of a set of hospital patients. (2) The entries of $\tilde{x}$ represent the length, width and height (all in centimeters) of a set of cars. (5 points)

If you compare the eigenvalues of $\Sigma_{\tilde{x}}$ there is a large difference between the last two eigenvalues and the first one and less differences between the eigenvalues of $\Sigma_{\tilde{y}}$. For the first case the variables have very different scales compared to the variables related to a set of cars, so it will make sense to use the directions of the PCA on $\Sigma_{\tilde{y}}$ to perform dimensionality reduction to 2D for the dataset related to hospital patients so one variable like height, does not dominate the two others. And we will do dimension reduction using the directions of PCA on $\Sigma_{\tilde{x}}$ for the data related to the length, width and height (all in centimeters) of a set of cars.

3. *Noise cancellation (*20 points*).* We are interested in recording the voice of a pilot in a helicopter. To this end we place a microphone inside his helmet and another microphone outside. We model the measurements as

$$\tilde{x}[1] = \tilde{y} + \alpha\tilde{z} \tag{4}$$
$$\tilde{x}[2] = \alpha\tilde{y} + \tilde{z}, \tag{5}$$

where $\tilde{y}$ is a random variable modeling the voice of the pilot, $\tilde{z}$ is a random variable modeling the noise in the helicopter, and $0 < \alpha < 1$ is a constant that models the effect of the helmet. From past data, we determine that $\tilde{y}$, and $\tilde{z}$ are zero mean and uncorrelated with each other. The variances of $\tilde{y}$ and $\tilde{z}$ are equal to 1 and 100 respectively.

a. Compute the best linear estimator of $\tilde{y}$ given $\tilde{x}[1]$ in terms of MSE, and the corresponding MSE, as a function of α. Describe in words what the estimator does. (10 points)

By independence

$$\begin{aligned}
\mathrm{Var}(\tilde{x}[1]) &= \mathrm{Var}(\tilde{y}) + \mathrm{Var}(\alpha\tilde{z}) \\
&= 1 + \alpha^2\mathrm{Var}(\tilde{z}) \\
&= 1 + 100\alpha^2 \\
\Sigma_{\tilde{x}[1]} &= 1 + 100\alpha^2 \\
\mathrm{Cov}(\tilde{y}(\tilde{x}[1])) &= \mathrm{E}[\tilde{y}(\tilde{x}[1])] \\
&= .\mathrm{Var}(\tilde{y}) \\
&= 1
\end{aligned}$$

By theorem 2.3 in linear regression notes, the estimate equals:

$$\hat{y}(\tilde{x}[1]) = \frac{\tilde{x}[1]}{100\alpha^2 + 1}$$

And the corresponding MSE equals:

$$\begin{aligned}
\mathrm{E}((\hat{y}(\tilde{x}[1]) - \tilde{y})^2] &= \mathrm{Var}(\tilde{y}) - \frac{1}{1 + 100\alpha^2} \\
&= 1 - \frac{1}{1 + 100\alpha^2} \\
&= \frac{100\alpha^2}{100\alpha^2 + 1}
\end{aligned}$$

If $\alpha = 0$ the estimate is the the voice of the pilot and when $\alpha \to 1$ then the estimate shrinks the voice of the pilot from the microphone. α controls how much shrinkage we want of the voice of the pilot.

b. Compute the best linear estimator of $\tilde{y}$ given $\tilde{x}$ in terms of MSE, and the corresponding MSE, as

a function of α. Describe in words what the estimator does. (10 points) By independence

$$\begin{aligned}
\mathrm{Var}(\tilde{x}[1]) &= 100\alpha^2 + 1 \\
\mathrm{Var}(\tilde{x}[2]) &= \mathrm{Var}(\alpha\tilde{y}) + \mathrm{Var}(\tilde{z}) \\
&= \alpha^2 + 100 \\
\mathrm{Cov}(\tilde{x}[1]\tilde{x}[2]) &= \mathrm{E}[(\tilde{y} + \alpha\tilde{z})(\alpha\tilde{y} + \tilde{z})] \\
&= \alpha E[\tilde{y}^2] + \alpha \mathrm{E}[\tilde{z})] \\
&= 101\alpha \\
\Sigma_{\tilde{x}} &= \begin{bmatrix} 100\alpha^2 + 1 & 101\alpha \\ 101\alpha & \alpha^2 + 100 \end{bmatrix} \\
\mathrm{Cov}(\tilde{y}(\tilde{x}[1])) &= 1 \\
\mathrm{Cov}(\tilde{y}(\tilde{x}[2])) &= \mathrm{E}[\tilde{y}(\alpha\tilde{y} + \tilde{z})] \\
&= \alpha \mathrm{E}[\tilde{y}^2] = \alpha \\
\Sigma_{\tilde{y}} &= \begin{bmatrix} 1 \\ \alpha \end{bmatrix}
\end{aligned}$$

By theorem 2.3 in the linear regression notes, the estimate equals:

$$\begin{aligned}
\hat{y}(\tilde{x}) &= \tilde{x}^T \begin{bmatrix} 100\alpha^2 + 1 & 101\alpha \\ 101\alpha & \alpha^2 + 100 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \\
&= \frac{\tilde{x}[1] - \tilde{x}[2]}{1 - \alpha^2}
\end{aligned}$$

And the corresponding MSE equals:

$$\begin{aligned}
\mathrm{E}((\hat{y}(\tilde{x}) - \tilde{y})^2] &= \mathrm{Var}(\tilde{y}) - \begin{bmatrix} 1 & \alpha \end{bmatrix} \begin{bmatrix} 100\alpha^2 + 1 & 101\alpha \\ 101\alpha & \alpha^2 + 100 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \\
&= 1 - \frac{\alpha - 1}{\alpha^2 - 1} \\
&= \frac{\alpha}{\alpha + 1}
\end{aligned}$$

The estimate as $\alpha \to 1$ approximatively cancels out the noise in a way that is optimal with respect to MSE, and provides better recording of the voice of the pilot.

4. *Adversarial perturbation (*30 points*).* Designing adversarial perturbations is an active area of research in deep learning. Here we study adversarial perturbations for linear regression.

   a. Let β_{OLS} be the OLS coefficient estimate corresponding to a linear-regression problem where $X \in \mathbb{R}^{p\times n}$ is the feature matrix and $y \in \mathbb{R}^n$ is the response vector. We assume that X is full rank and $p < n$. Your goal is to choose a noise vector $z \in \mathbb{R}^n$ with bounded ℓ_2 norm ($\|z\|_2 \leq \gamma$ for a constant $\gamma > 0$) that will be added to y. Explain how to compute the z that produces the maximum possible perturbation (in ℓ_2 norm) to the OLS coefficient estimate, i.e.

$$z^{\text{adv}}_{\text{OLS}} := \arg\max_{||z||_2\leq\gamma} \left|\left|\beta_{\text{OLS}} - \beta^{\text{mod}}_{\text{OLS}}(z)\right|\right|_2^2, \quad \beta^{\text{mod}}_{\text{OLS}}(z) := \arg\min_{\beta} \left|\left|y + z - X^T\beta\right|\right|_2^2. \tag{6}$$

   Using theorem 2.4 from linear regression notes, we have

$$\beta^{\text{mod}}_{\text{OLS}}(z) = (XX^T)^{-1}X(y+z)$$

   Let $X = USV^T$ then

$$\begin{aligned}
\beta^{\text{mod}}_{\text{OLS}}(z) &= (USV^TVSU^T)^{-1}USV^T(y+z)\\
&= US^{-2}U^TUSV^T(y+z)\\
&= US^{-1}V^T(y+z)\\
\beta_{\text{OLS}} - \beta^{\text{mod}}_{\text{OLS}} &= US^{-1}V^Tz\\
&= \sum_{i=1}^{p}\frac{v_i^Tz}{s_i}u_i\\
\arg\max_{||z||_2\leq\gamma} \left|\left|\beta_{\text{OLS}} - \beta^{\text{mod}}_{\text{OLS}}(z)\right|\right|_2^2 &= \arg\max_{||z||_2\leq\gamma} \|US^{-1}V^Tz\|_2^2\\
&= \arg\max_{||z||_2\leq\gamma} z^TVS^{-2}V^Tz
\end{aligned}$$

   By theorem 3.1 from linear regression notes, the optimal noise vector is the vector corresponding to the last row of V, $z = V[:, p]$, which matches the largest singular value of S^{-1} which is $\frac{1}{s_n}$ where $s_n \leq \ldots \leq s_1$ are the singular values of X. z being the last right-singular vector has unit norm. With this value of $z^{\text{adv}}_{\text{OLS}}$, $\left|\left|\beta_{\text{OLS}} - \beta^{\text{mod}}_{\text{OLS}}(z^{\text{adv}}_{\text{OLS}})\right|\right|_2^2 = \frac{1}{s_n}$.

   (10 points)

   b. Let $\tilde{x}_{\text{test}}$ be a random test feature vector. Consider the perturbation to the test prediction $x^T_{\text{test}}\beta(z_{\text{adv}}) - x^T_{\text{test}}\beta_{\text{OLS}}$ achieved by your choice in the previous question. Do you expect the variance of this perturbation to be larger if $\frac{1}{n}XX^T$ is a good approximation to the covariance matrix of $\tilde{x}_{\text{test}}$, or if it is not a good approximation? Why? (5 points)
   Using the theorem 4.8 from the PCA notes, we find

$$\text{Var}(\tilde{x}^T_{\text{test}}(\beta(z_{\text{adv}}) - \beta_{\text{OLS}})) = \sum_{i=1}^{p}\frac{u_i^T\Sigma_{\tilde{x}_{\text{test}}}u_i}{s_i^2}$$

   If this sample variance is a good approximation to the variance of the test data in that direction then

$$\text{Var}(\tilde{x}^T_{\text{test}}(\beta(z_{\text{adv}}) - \beta_{\text{OLS}})) \approx \frac{p}{n}$$

If it is not a good approximation then the sample covariance matrix may not provide a good estimate of the feature variance in every direction, in that case, there may be terms in the variance where s_i is very small, due to correlations between the features, the variance of this perturbation will be larger.

c. Assume that the regularization parameter $\lambda > 0$ is known to you. Explain how to choose the z that produces the maximum possible perturbation (in ℓ_2 norm) to the ridge regression coefficient estimate, i.e.

$$z_{\text{RR}}^{\text{adv}}(\lambda) := \arg\max_{||z||_2 \leq \gamma} \left|\left|\beta_{\text{RR}}(\lambda) - \beta_{\text{RR}}^{\text{mod}}(z,\lambda)\right|\right|_2^2, \quad \beta_{\text{RR}}^{\text{mod}}(z,\lambda) := \arg\min_{\beta} \left|\left|y + z - X^T\beta\right|\right|_2^2 + \lambda\,||\beta||_2^2\,.$$

Is the perturbation $\left|\left|\beta_{\text{RR}}(\lambda) - \beta_{\text{RR}}^{\text{mod}}(z_{\text{RR}}^{\text{adv}}(\lambda),\lambda)\right|\right|_2^2$ larger or smaller than the one achieved for OLS in the previous question, $\left|\left|\beta_{\text{OLS}} - \beta_{\text{OLS}}^{\text{mod}}(z_{\text{OLS}}^{\text{adv}})\right|\right|_2^2$? (10 points)

Using theorem 5.2 from linear regression notes, we have

$$\begin{aligned}
\beta_{\text{RR}}(\lambda) &= (XX^T + \lambda I)^{-1}Xy \\
\beta_{\text{RR}}^{\text{mod}}(z,\lambda) &= (XX^T + \lambda I)^{-1}X(y+z) \\
\beta_{\text{RR}}(\lambda) - \beta_{\text{RR}}^{\text{mod}}(z,\lambda) &= (XX^T + \lambda I)^{-1}Xz \\
&= (USV^TVSU^T + \lambda UU^T)^{-1}Xz \\
&= U(S^2 + \lambda I)^{-1}U^TUSV^Tz \\
&= U(S^2 + \lambda I)^{-1}SV^Tz \\
&= \sum_{i=1}^{p} \frac{s_i \langle v_i, z\rangle}{s_i^2 + \lambda} u_i \\
\arg\max_{||z||_2 \leq \gamma} \left|\left|\beta_{\text{RR}}(\lambda) - \beta_{\text{RR}}^{\text{mod}}(z,\lambda)\right|\right|_2^2 &= \arg\max_{||z||_2 \leq \gamma} |||U(S^2 + \lambda I)^{-1}SV^Tz||_2^2 \\
&= \arg\max_{||z||_2 \leq \gamma} z^TV^T(S^2 + \lambda I)^{-2}S^2V^Tz
\end{aligned}$$

For a fixed $\lambda > 0$, z will be the right singular vector v_i, corresponding to the largest entry $\frac{s_i}{s_i^2+\lambda}$. Setting $\lambda = 0$ we have the same perturbation as the perturbation obtained in part a, and as $\lambda \to \infty$, the maximum perturbation to the ridge regression coefficient estimate goes to 0. The difference between the two perturbations is $\frac{1}{s_n} - \frac{s_i}{s_i^2+\lambda}$ which could be positive or negative depending the values of the singular values.

d. How would you set λ to minimize the ℓ_2 norm of the adversarial perturbation, $\left|\left|\beta_{\text{RR}}(\lambda) - \beta_{\text{RR}}^{\text{mod}}(z_{\text{RR}}^{\text{adv}}(\lambda),\lambda)\right|\right|_2^2$? Is this a reasonable approach? (5 points)

The perturbation is minimized by increasing λ. Increasing λ reduces the adversarial perturbation, but as described in the PCA notes in the chapter about ridge regression, too large values for λ reduces the ridge regression coefficients estimates, which become eventually too small to produce an accurate fit. There is a trade-off to achieve and it is determined using cross-validation.