CS 189 Introduction to Machine Learning
Fall 2017 EXAMMIDTERM

1 Multiple Choice Questions (14 points)

For these questions, select *all* the answers which are correct. You will get full credit for selecting all the right answers. On some questions, partial credit will be assigned.

(a) Increasing λ in ridge regression can be interpreted as performing a MAP estimate with a

◯ Gaussian prior with smaller variance
◯ Uniform prior with smaller range
◯ Gaussian prior with larger variance
◯ Uniform prior with larger range

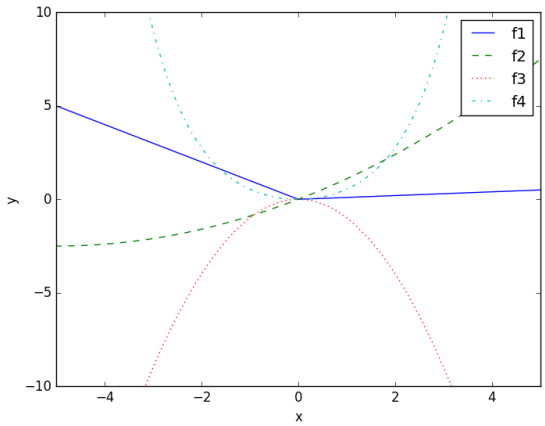
Solution: The correct answer is **Gaussian prior with smaller variance.**

(b) How does total least squares (TLS) compare to ordinary least squares (OLS)?

◯ TLS allows errors in X and y. OLS only allows errors in X.
◯ TLS only allows errors in X. OLS allows errors in X and y.
◯ TLS allows errors in X and y. OLS only allows errors in y.
◯ TLS only allows errors in y. OLS allows errors in X and y.

Solution: The correct answer is **TLS allows errors in X and y. OLS only allows errors in y.**

(c) (NOT IN SCOPE FOR SP18 EXAM) Which of the following functions are convex? They are drawn in the following plot and defined explicitly below:



◯ $f_1(x) = \max\{-x, 0.1x\}$

◯ $f_2(x) = x + \frac{x^2}{10}$

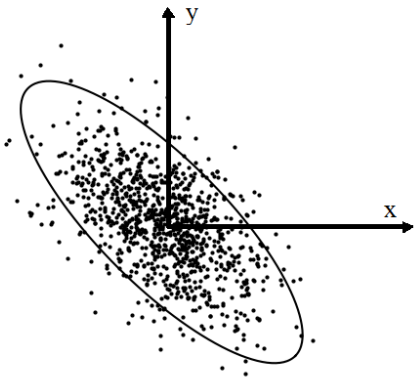
◯ $f_3(x) = -x^2$

◯ $f_4(x) = \frac{e^x + e^{-x}}{2} - 1$

Solution: The correct answers are f_1, f_2, f_4.

f_1 is convex because it is the max of convex functions. f_2 is convex because it is the sum of convex functions. f_4 is convex because it is the sum of convex functions multiplied by a constant $(\frac{1}{2})$.

(d) Select which covariance matrix was most likely used to generate the following multivariate Gaussian distribution



where the positive x direction is to the right and the positive y direction is up.

◯ $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$

◯ $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

◯ $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

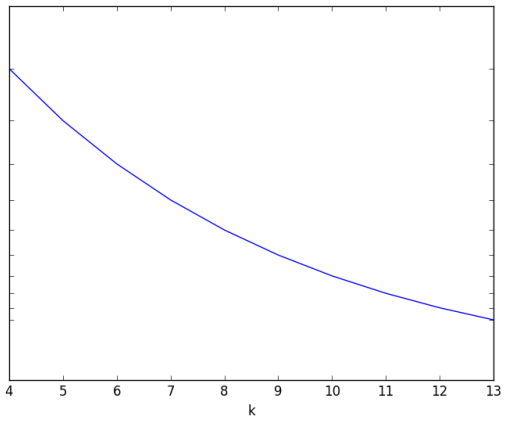
◯ $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

Solution: The correct answer is $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$.

There are two key observations. First, x and y are *negatively correlated* because when x increases, y decreases. This means the solution is either $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ or $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$. Second, $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ is a *degenerate* Gaussian because it has an eigenvalue of 0, so there would be no variation in the direction $\vec{d} = [1,1]^T$.

(e) Your friend at Google is training a machine learning model to predict a user's next search query based on their past k searches. She generates the following plot, where the value of k

is on the x axis.



What might the y axis represent?

◯ Training Error ◯ Bias

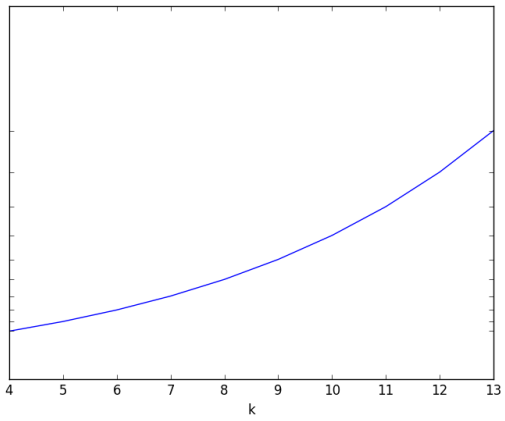
◯ Validation Error ◯ Variance

Solution: As k increases, the number of features in the ML model increases.

Full credit was given to anyone who responded **Training Error, Bias, Validation Error** or **Training Error, Bias**.

It is most important to recognize that training error and bias decrease with the number of features in the ML model. The reason that validation error is a correct answer is because it may be the case that the behavior changes after $k = 13$. Given this graph, it is not possible to know for sure.

(f) Your friend at Google is training a machine learning model to predict a user's next search query based on their past k searches. She generates the following plot, where the value of k is on the x axis.



What might the y axis represent?

◯ Training Error ◯ Bias

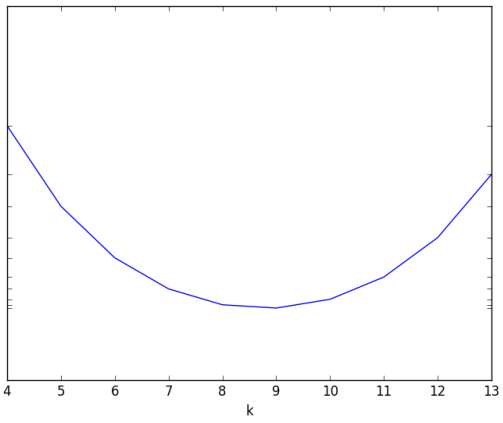
◯ Validation Error ◯ Variance

Solution: As k increases, the number of features in the ML model increases.

Full credit was given to anyone who responded **Variance, Validation Error** or **Variance**.

It is most important to recognize that variance increases with the number of features in the ML model. The reason that validation error is a correct answer is because it may be the case that the behavior changes before $k = 4$. Given this graph, it is not possible to know for sure.

(g) Your friend at Google is training a machine learning model to predict a user's next search query based on their past k searches. She generates the following plot, where the value of k is on the x axis.



What might the y axis represent?

◯ Training Error ◯ Bias

◯ Validation Error ◯ Variance

Solution: As k increases, the number of features in the ML model increases.

Full credit was given to anyone who responded **Validation Error**.

2 Short Answer Questions (7 points)

For these questions, you only need to give an answer. You do not need to show your work. You will only be judged on the correctness of your answer.

(a) (3 points) Suppose we have a covariance matrix for zero-mean X:

$$E[XX^T] = \Sigma = \begin{bmatrix} 4 & a \\ a & 1 \end{bmatrix}$$

What is the set of values that a can take on such that Σ is a valid covariance matrix?

Solution: The matrix Σ must be PSD, meaning both its eigenvalues must be non-negative. We can calculate the eigenvalues by calculating

$$(4-\lambda)(1-\lambda) - a^2 = 0 \implies \lambda^2 - 5\lambda + (4-a^2) = 0 \implies \lambda = \frac{5 \pm \sqrt{25 - 4(4-a^2)}}{2}.$$

Now,

$$\lambda \geq 0 \implies 5 - \sqrt{25 - 16 + 4a^2} \geq 0 \implies 25 \geq 9 + 4a^2 \implies 4 \geq a^2.$$

The correct answer is $|a| \leq 2$. We also gave full credit for answering $|a| < 2$, since $|a| = 2$ gives a degenerate Gaussian (which is still technically a Gaussian).

(b) (4 points) Suppose that we observe the position and velocity of an object moving **along a line** in 3D space. At any point on the line, the object can have any speed. Our position observations measure the x, y, and z coordinates of the object, and the velocity observations measure the x, y, and z components of the velocity. We collect a large set of observations and run PCA on the set. **How many principal components would we expect to use to represent this data set?**

Solution: The intended solution was 2. You need one parameter to describe the location of the object and one parameter to describe the velocity at that location (since the velocity is along the line, we only care about its magnitude and not its direction, so we only care about the speed).

However, the solution 2 assumes that the line goes through the origin. If the line does not go through the origin, then you cannot describe the position of the object with a single parameter, so the answer is 4: three parameters for the location and still one parameter for the velocity (again, we only care about speed, i.e. the magnitude of the velocity, since the direction is always along the line).

Hopefully thinking about this problem helps you connect the idea of CCA and the idea of "degrees of freedom."

3 Parameter Estimation (19 points)

Assume that $X_1, X_2, ..., X_n$ are i.i.d. samples from a Poisson distribution:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where $\lambda > 0$ and x is a non negative-integer. $P(X < 0) = 0$ and $P(X = x) = 0$ for all non-integer x.

(a) (3 points) **Compute the log likelihood of drawing samples $X_1 = x_1, \ldots, X_n = x_n$ for a given λ, i.e., $\ln p(x_1, x_2, \ldots, x_n|\lambda)$.**

Solution:

$$l(\lambda) = \log(\Pi_{i=1}^n P(X_i|\lambda)) = \log(\Pi_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}) = \Sigma_{i=1}^n X_i \log\lambda - \lambda - \log(X_i!)$$
$$= (\Sigma_{i=1}^n X_i)\log\lambda - n\lambda - \Sigma_{i=1}^n \log(X_i!)$$

(b) (8 points) **Show that the negative of the log likelihood above is a convex function with respect to λ.** Use this fact to **compute the maximum likelihood estimate of λ given samples $X_1 = x_1, \ldots, X_n = x_n$.**

Solution:

Negative log likelihood is a convex function. Observe that $-\log\lambda$ is convex because the second derivative is $1/\lambda^2$ and it is always non negative. Also, $n\lambda + \Sigma_{i=1}^n \log(X_i!)$ is convex because it is a linear function of λ. Because sum of convex functions is convex, negative log likelihood is convex. We can compute the MLE by setting the derivative of log likelihood with respect to λ to zero:

$$\frac{dl(\lambda)}{d\lambda} = \frac{1}{\lambda}\Sigma_{i=1}^n X_i - n = 0$$
$$\lambda = \frac{\Sigma_{i=1}^n X_i}{n}$$

(c) (8 points) Now assume that we have a prior on λ with an exponential density $f(\lambda) = \alpha e^{-\alpha\lambda}$. **Compute the maximum a posteriori estimate of λ given samples $X_1 = x_1, \ldots, X_n = x_n$. Compare the MAP and MLE. What happens when $n \to \infty$?**

Solution:

$$P(\lambda|X_1, ..., X_n) \propto P(X_1, ..., X_n|\lambda)P(\lambda) \propto (\Pi_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!})\alpha e^{-\alpha\lambda}$$

$$\propto (\Pi_{i=1}^n \lambda^{X_i} e^{-\lambda}) e^{-\alpha\lambda}$$

$$l(\lambda) = \Sigma_{i=1}^{n} X_i \log \lambda - \lambda n - \alpha\lambda$$

$$\frac{dl(\lambda)}{d\lambda} = 0$$

$$\frac{\Sigma_{i=1}^{n} X_i}{\lambda} - n - \alpha = 0$$

$$\lambda = \frac{\Sigma_{i=1}^{n} X_i}{n + \alpha}$$

If n is large the MAP and MLE estimation will be the same. In general, if we don't know the prior on parameters, we use MLE instead of MAP. Having a good prior on parameters helps to estimate parameters more accurately when the sample size is small. However, if we don't know the prior on parameters, having a large dataset will compensate for that.

4 Linear regression (16 points)

In this problem, we study the loss function for ridge regression:

$$\frac{1}{2}\|Xw-y\|_2^2+\frac{\lambda}{2}\|w\|_2^2, \tag{1}$$

where X a data matrix in $\mathbb{R}^{n\times d}$ and y is the response in $\mathbb{R}^n$. Let the regularization weight $\lambda > 0$. Recall that the closed-form solution to ridge regression is

$$\hat{w} = (X^TX+\lambda I_d)^{-1}X^Ty. \tag{2}$$

where $I_d \in \mathbb{R}^{d\times d}$ is an identity matrix of dimension d.

(a) (8 points) Augment the matrix X with d additional rows $ce_1^T, ce_2^T, \ldots, ce_d^T$ to get the matrix $X' \in \mathbb{R}^{n+d\times d}$, where c is a given constant and e_i^T is a unit vector whose ith element is 1 and the rest of the elements are zero, and augment y with d zeros to get $y' \in \mathbb{R}^{n+d}$:

$$X' = \begin{bmatrix} X \\ ce_1^T \\ ce_2^T \\ \vdots \\ ce_d^T \end{bmatrix} \qquad y' = \begin{bmatrix} y \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Write out the closed form solution for w' for the ordinary least squares problem on X', y'.

Derive the concrete value of c that corresponds to the ridge regression problem (1) **in terms of λ.** Conclude that the ridge regression estimate for w can be obtained by doing ordinary least squares on an augmented dataset.

Solution: Augmenting the matrix X with d additional rows $ce_1^T, ce_2^T, \ldots, ce_d^T$, we get a new matrix X' such that

$$X' = \begin{bmatrix} X \\ ce_1^T \\ ce_2^T \\ \vdots \\ ce_d^T \end{bmatrix} \tag{3}$$

Augmenting y with d zeros, we get a new vector y' such that

$$y' = \begin{bmatrix} y \\ 0_d \end{bmatrix} \tag{4}$$

where 0_d is an all-zero vector in $\mathbb{R}^d$. Solving the least square problem with X', y', we have

$$\begin{aligned} w' &= (X'^T X')^{-1} X' y' \\ &= \left(\begin{bmatrix} X^T & ce_1 & ce_2 & \dots & ce_d \end{bmatrix} \begin{bmatrix} X \\ ce_1^T \\ ce_2^T \\ \vdots \\ ce_d^T \end{bmatrix}\right)^{-1} \begin{bmatrix} X^T & ce_1 & ce_2 & \dots & ce_d \end{bmatrix} \begin{bmatrix} y \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= (X^T X + c^2 \sum_{i=1}^{d} e_i e_i^T)^{-1} X^T y \\ &= (X^T X + c^2 I_d)^{-1} X^T y. \end{aligned}$$

The closed form solution of the ridge regression is

$$w' = (X^T X + \lambda I_d)^{-1} X^T y. \tag{5}$$

Therefore, we have $\lambda = c^2$. So $c = \sqrt{\lambda}$ or $c = -\sqrt{\lambda}$.

(b) (8 points) **(NOT IN SCOPE FOR SP18 MIDTERM) Prove that applying gradient descent on the ridge regression loss function with a suitable fixed step size results in geometric convergence.** If $X^\top X$ has maximum and minimum eigenvalues M and m, **what fixed step size should we choose as a function of M, m, and ridge weight λ?**

Hint: Reduce the problem to ordinary least squares. Recall that applying gradient descent on the least squares problem

$$\min_x f(x) = \min_x \frac{1}{2}\|Ax - b\|_2^2$$

results in geometric convergence when $A^T A$ is positive definite and using a constant step size $\gamma = \frac{2}{\lambda_{\min}(A^T A) + \lambda_{\max}(A^T A)}$. Geometric convergence means that $f(x_k) - f(x^*) \leq c' Q^k$ for some $0 \leq Q < 1$ and for some $c' > 0$. You may use this result without proof.

Solution: Let m_1, m_2 denote the smallest and largest eigenvalues of $X^T X$, respectively. We can formulate the problem as

$$\min_x \frac{1}{2}\|A'x - b'\|_2^2$$

where

$$A' = \begin{bmatrix} X \\ \lambda I_d \end{bmatrix} \text{ and } b' = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

We observe that $A'^T A' = X^T X + \lambda I_d \succeq m_1 + \lambda > 0$ (positive definite). That is, we reduce the given problem to the quadratic problem above. The smallest and largest eigenvalues of $X^T X + \lambda I_d$ are $m_1 + \lambda, m_2 + \lambda$, respectively. Then, by the given result, geometric convergence is attained for fixed step size $\gamma = \frac{2}{m_1 + m_2 + 2\lambda}$.

5 Finding noisy low-rank matrices (21 points)

Assume you have an *unknown* matrix $M^* \in \mathscr{L}$, where $\mathscr{L}$ represents the set of all $d \times d$ square matrices of rank up to k. You observe M^* through noise as $Y = M^* + N$, where $N \in \mathbb{R}^{d \times d}$ is a noise matrix with $N_{ij} \overset{i.i.d.}{\sim} \mathscr{N}(0,1)$.

(a) (8 points) **Show that the maximum likelihood estimate of the matrix M^* is given by the optimization problem**

$$\widehat{M} = \arg\min_{M \in \mathscr{L}} \|Y - M\|_F^2, \tag{6}$$

where $\|X\|_F^2 = \sum_{i,j} X_{ij}^2$ is the squared Frobenius norm.

Hint: Begin by writing the likelihood $P(Y|M)$ for some matrix M. The exact definition of $\mathscr{L}$ is not relevant to this part; we will use this in the next part.

Solution: Note that here, we don't know the underlying matrix M, and want to estimate it given the data Y. Note that the entry i, j of Y is distributed as $Y_{ij} \sim \mathscr{N}(M_{ij}, 1)$. The likelihood is therefore given by

$$\begin{aligned} P(Y|M) &= \prod_{i,j} P(Y_{ij}|M_{ij}) \\ &\propto \exp\left(-\frac{1}{2}(Y_{ij} - M_{ij})^2\right) \\ &= \exp\left(-\frac{1}{2}\sum_{i,j}(Y_{ij} - M_{ij})^2\right). \end{aligned}$$

Our goal is to maximize this likelihood, so it is equivalent to maximize the log-likelihood. Writing this out using the definition of Frobenius norm and using c to denote a constant, we have

$$\log P(Y|M) = -\|Y - M\|_F^2 + c.$$

Maximizing the log-likelihood over $M \in \mathscr{L}$ is therefore equivalent to performing the minimization $\min_{M \in \mathscr{L}} \|Y - M\|_F^2$.

(b) (8 points) Recall that $\mathscr{L}$ was the class of $d \times d$ matrices of rank at most k. Assume that Y is full rank (i.e. invertible). Let us now consider the equivalent optimization problem

$$\widehat{M} = \arg\min_{M:\mathsf{rank}(M)=k} \|Y - M\|_F^2. \tag{7}$$

Write down a closed form expression to solve this problem when $k = d - 1$ in terms of the singular values and singular vectors of the matrix Y.

Hint: Your knowledge of solving total least squares problems may come in handy. (No need to derive the solution, if you can justify it with TLS.)

Solution: There is a direct way to approach this problem based on a theorem we know from class/HW called the Eckart-Young theorem. In particular, for a matrix Y having singular value decomposition $Y = \sum_{i=1}^{d} \sigma_i u_i v_i^\top$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ (since Y is full rank), the theorem states that the rank k matrix M minimizing $\|Y - M\|_F^2$ is given by $\hat{M} = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$. Specializing to the case $k = d-1$ yields $\hat{M} = \sum_{i=1}^{d-1} \sigma_i u_i v_i^\top$.

Alternatively, we could have used the process of TLS to arrive at a similar solution. Assuming that we have an $m \times n$ matrix A with $m > n$, we have a rank $n+1$ matrix $[A|y]$, and would like to find the minimum (Frobenius) norm perturbation $[\hat{A}|\hat{y}]$ such that $[A+\hat{A}|y+\hat{y}]$ has rank n. Let us assume that $m = n+1$. Then we are solving the problem $\min_{M \in \mathscr{K}} \|[A|y] - M\|_F^2$, where $\mathscr{K}$ is the space of square matrices of rank n. We know that the solution to this problem is given by truncating the last singular value of the matrix $[A|y]$, and setting $n = d-1$ recovers our setting and result.

(c) (5 points) **BONUS: Write down a closed form expression to the problem above for arbitrary k in terms of the singular values and singular vectors of the matrix Y.**

Solution: As mentioned above, this comes stright from the Eckart-Young theorem. If Y has SVD $Y = \sum_{i=1}^{d} \sigma_i u_i v_i^\top$, then the *rank-k approximation* is given by $\hat{M} = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$.

6 (NOT IN SCOPE FOR SP18 MIDTERM) Multi-view Regression with CCA (30 points)

In robotics and other problem domains, simulations are a cheap source of training data. However, simulated data isn't perfect and might not present things in the same way that real data does. In this problem, we will show how CCA can help us use simulated data to learn by allowing us to focus on those dimensions of the simulation that actually correspond to the real world.

Let $X \in \mathbb{R}^d$ be a zero-mean random variable representing simulated data. Let $Q \in \mathbb{R}^d$ be a zero-mean random variable representing real-world data. Our goal is to learn the output $Y \in \mathbb{R}$ which is some bounded control signal (i.e. $Y \in [-1,1]$).

For example, one view Q could be images of an object taken from the real world and the other view X could be a corresponding image from a simulator. While the views are different we expect the robot to grasp (using control Y) the object in the same way.

Suppose that we know the simulation/real-world correspondences via their covariance matrix $\mathbb{E}[XQ^T] = \Sigma_{XQ}$. We also know the individual positive-definite covariance matrices Σ_{XX} and Σ_{QQ}.

(a) (8 points) In order to leverage CCA, we compute the canonical variates of our data and transform the data it into a space where the views are most correlated. Denote the canonical correlation matrix

$$C = \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XQ} \Sigma_{QQ}^{-\frac{1}{2}}.$$

Let $C = U\Lambda V^T$ be its singular value decomposition. We can transform our data to the canonical variates via the following:

$$\hat{X} = U^T \Sigma_{XX}^{-\frac{1}{2}} X \qquad\qquad \hat{Q} = V^T \Sigma_{QQ}^{-\frac{1}{2}} Q$$

Show that $\mathbb{E}[\hat{X}\hat{Q}^T] = \Sigma_{\hat{X}\hat{Q}} = \Lambda$ and $\Sigma_{\hat{X}\hat{X}} = \Sigma_{\hat{Q}\hat{Q}} = I$.

Here, we use the symmetric square root: For a positive definite matrix Σ having eigenvalue decomposition $\Sigma = \bar{U}\bar{\Lambda}\bar{U}^\top$, the symmetric square-root is given by $\Sigma^{\frac{1}{2}} = \bar{U}\bar{\Lambda}^{\frac{1}{2}}\bar{U}^T$.

Solution: Given the whitened and decorrelated simulated data $\hat{X}$, we can compute the covariance matrix as follows:

$$\begin{aligned}
\Sigma_{\hat{X}\hat{X}} &= \mathbb{E}[\hat{X}(\hat{X})^T] \\
&= \mathbb{E}[(U^T \Sigma_{XX}^{-\frac{1}{2}} X)(U^T \Sigma_{XX}^{-\frac{1}{2}} X)^T] \\
&= (U^T \Sigma_{XX}^{-\frac{1}{2}} \mathbb{E}[XX^T] \Sigma_{XX}^{-\frac{1}{2}} U \\
&= U^T \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XX} (\Sigma_{XX}^{-\frac{1}{2}})^T U \\
&= U^T I U \\
&= I,
\end{aligned}$$

where have used linearity of expectation to bring the expectation into the expression, and the fact that $\Sigma_{XX}^{-1/2}$ is symmetric.

The same can be shown for $\hat{Q}$. We next look at the cross-covariance matrix.

$$\begin{aligned}
\Sigma_{\hat{X}\hat{Q}} &= \mathbb{E}[\hat{X}\hat{Q}^T] \\
&= \mathbb{E}[(U^T\Sigma_{XX}^{-\frac{1}{2}}X)(V^T\Sigma_{QQ}^{-\frac{1}{2}}Q)^T] \\
&= \mathbb{E}[(U^T\Sigma_{XX}^{-\frac{1}{2}}XQ^T\Sigma_{QQ}^{-\frac{1}{2}}V] \\
&= (U^T\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XQ}\Sigma_{QQ}^{-\frac{1}{2}}V \\
&= U^T(U\Lambda V^T)V \\
&= \Lambda.
\end{aligned}$$

(b) (4 points) To focus attention on the dimensions of the simulation that most correspond to the real world, we can use regularization while we attempt to learn the best linear map from simulated X to control Y. We define

$$\hat{w} = \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \ \mathbb{E}[(Y - w^T\hat{X})^2] + ||w||^2_{\text{CCA}} \tag{8}$$

where

$$||w||^2_{CCA} = \sum_{i=1}^{d} \frac{1-\lambda_i}{\lambda_i}(w_i)^2.$$

Here, the λ_i correspond to the diagonal elements of Λ, and satisfy $0 < \lambda_i \leq 1$ by the properties of CCA. The weighted norm here penalizes dependencies on those aspects of simulated X that are less correlated to the real Q. The expectation is taken over both random variables $\hat{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$.

Show the following is true

$$\mathbb{E}[(Y - w^T\hat{X})^2] = \mathbb{E}[Y^2] + \|w\|_2^2 - 2w^T\mathbb{E}[Y\hat{X}].$$

Solution:

We have

$$\begin{aligned}
\mathbb{E}(Y - w^T\hat{X})^2 &= \mathbb{E}[Y^2 - 2w^T\hat{X}Y + (w^T\hat{X})(w^T\hat{X})] \\
&= \mathbb{E}[Y^2] - 2w^T\mathbb{E}[\hat{X}Y] + w^T\mathbb{E}[\hat{X}\hat{X}^T]w \\
&= \mathbb{E}[Y^2] - 2w^T\mathbb{E}[\hat{X}Y] + \|w\|_2^2
\end{aligned}$$

Where the last fact follows from the covariance matrix being identity, which was proved in the last part.

(c) (8 points) **Show the global minimizer $\hat{w}$ in Eq. (8) has coordinate i given by**

$$\hat{w}_i = \lambda_i \mathbb{E}[Y(\hat{X})_i],$$

where $(\hat{X})_i$ is the i-th coordinate of $\hat{X}$.

Solution:

To determine the solution to the problem, we will compute the gradient of the expectation and set it equal to zero. (Note that this is sufficient since the problem is convex, but you weren't expected to show that for full credit.)

Using the previous part, we have

$$0 = \nabla_w \mathbb{E}[Y^2] + ||w||_2^2 - 2w^T \mathbb{E}[\hat{X}Y] + ||w||_{\text{CCA}}^2.$$

Notice that each coordinate i appears separately in the sum above, so we can take the derivative with respect to w_i separately and set it to zero. Consequently, we have

$$2w_i - 2w_i \mathbb{E}[\hat{X}_i Y] + 2\frac{1-\lambda_i}{\lambda_i} w_i = 0.$$

Solving for the ith coodinate (which is given by $\hat{w}_i$), we have

$$\hat{w}_i = \lambda_i \mathbb{E}[Y\hat{X}_i].$$

(d) (10 points) Let us take n samples $(x^1, y^1), \ldots, (x^n, y^n)$ of the view X and the corresponding outputs Y, where each sample $x^j \in \mathbb{R}^d$ and $y^j \in \mathbb{R}$. We now transform the data to the canonical variates to obtain $\hat{x}^j = U^T \Sigma_{XX}^{-\frac{1}{2}} x^j$. We now use the data to create empirical estimates for $\hat{w}_i = \lambda_i \mathbb{E}[Y(\hat{X})_i]$ by defining

$$\tilde{w}_i = \lambda_i \frac{1}{n} \sum_{j=1}^{n} y^j (\hat{x}^j)_i$$

for each $i \in \{1, 2, \ldots, d\}$. As before, $(\hat{x}^j)_i$ denotes the ith coordinate of $\hat{x}^j$.

Let us examine how fast $\tilde{w}$ converges to $\hat{w}$. Notice that $\mathbb{E}[\tilde{w}] = \hat{w}$, therefore

$$\mathbb{E}[||\tilde{w} - \hat{w}||_2^2] = \underbrace{\mathbb{E}||\tilde{w} - \mathbb{E}(\tilde{w})||_2^2}_{\text{Variance}}$$

Show that

$$\mathbb{E}[||\tilde{w} - \hat{w}||_2^2] \leq \sum_{i=1}^{d} \frac{\lambda_i^2}{n}.$$

Hint: Remember the random variable $Y \in [-1,1]$ and so $Y^2 \leq 1$.

Solution:

From the description, it is sufficient to bound the quantity

$$\mathbb{E}[\|\tilde{w} - \mathbb{E}(\tilde{w})\|_2^2] = \sum_{i=1}^{d} \mathbb{E}[(\tilde{w}_i - \mathbb{E}[\tilde{w}_i])^2] = \sum_{i=1}^{d} \text{var}(\tilde{w}_i).$$

Let us now look at the variance of a particular coordinate. We have

$$\text{var}(\tilde{w}_i) = \text{var}\left(\lambda_i \frac{1}{n} \sum_{j=1}^{n} y^j (\hat{x}^j)_i\right),$$

and since each sample is drawn i.i.d., this is the variance of the sum of independent random variables. Using the facts that $\text{var}(aU) = a^2 \text{var}(U)$, and $\text{var}(U+V) = \text{var}(U) + \text{var}(V)$ for independent random variables U, V and a scalar a, we have

$$\text{var}(\tilde{w}_i) = \frac{\lambda_i^2}{n^2} n \,\text{var}\left(y^1 (\hat{x}^1)_i\right),$$

where we see that it is sufficient to look at the variance of our first sample. Note that this is equal to the variance of the random variable itself, so we have

$$\begin{aligned} \text{var}(\tilde{w}_i) &= \frac{\lambda_i^2}{n} \text{var}\left(Y (\hat{X})_i\right) \\ &\leq \frac{\lambda_i^2}{n} \mathbb{E}[Y^2 (\hat{X})_i^2], \end{aligned}$$

where we have used the fact that

$$\text{var}(U) = \mathbb{E}[U^2] - (\mathbb{E}[U])^2 \leq \mathbb{E}[U^2].$$

We can now use the fact $Y \in [-1,1]$ to see that $Y^2 (\hat{X})_i^2 \leq (\hat{X})_i^2$, and so $\mathbb{E}[Y^2 (\hat{X})_i^2] \leq \mathbb{E}[(\hat{X})_i^2]$, and so we have

$$\text{var}(\tilde{w}_i) \leq \frac{\lambda_i^2}{n} \mathbb{E}[Y^2 (\hat{X})_i^2] = \frac{\lambda_i^2}{n},$$

since $\hat{X}_i$ is whitened and zero mean. Summing over i in the given range completes the argument.

This result has a nice interpretation in relation to traditional OLS. Normally, the variance would be expected to scale at a rate of $\frac{d}{n}$, however now it scales by $\sum_{i=1}^{d} \lambda_i$ instead of d. Thus, if the views do not contain strong correlation, it suggests less data is needed to learn the function.

Intuitively, if a function can be learned on two views and produce the same output. Then it implies only the "intersection" of the views is needed to learn the data. Thus, the effective dimension of the data is smaller than the actual dimension. The augmented ridge regression, we consider implicitly finds this intersection. In Homework 5, we examined the hard projection case when doing Mooney Reconstruction.

In relation to leveraging simulation data, it suggests that simulated data can be used to reduce the effective dimension of the real world data. For example, to predict how to grasp an object a robot only needs relatively crude information from the world. However, natural RGB images are very rich in data. By leveraging simulation data and this CCA regularization term, we can reduce the effective dimension of the RGB image.