CS 189 / 289 Introduction to Machine Learning
Fall 2022 Jennifer Listgarten, Jitendra Malik HW1

Due 09/20/22 11:59 pm PT

- Homework 1 consists of both written and coding questions.
- We prefer that you typeset your answers using LaTeX or other word processing software. If you haven't yet learned LaTeX, one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted for the written questions.
- In all of the questions, **show your work**, not just the final answer.

Deliverables:

Submit a PDF of your homework to the Gradescope assignment entitled "HW1 Write-Up". **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.

- In your write-up, please state with whom you worked on the homework. If you worked by yourself, state you worked by yourself. This should be on its own page and should be the first page that you submit.
- In your write-up, please copy the following statement and sign your signature underneath. If you are using LaTeX, you must type your full name underneath instead. We want to make it *extra* clear so that no one inadvertently cheats. *"I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."*
- **Replicate all your code in an appendix**. Begin code for each coding question in a fresh page. Do not put code from multiple questions in the same page. When you upload this PDF on Gradescope, *make sure* that you assign the relevant pages of your code from appendix to correct questions.

1 (12 points) Multivariate Gaussians: A review

(a) (4 points) Consider a two dimensional, zero mean random variable $Z = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix}^\top \in \mathbb{R}^2$. In order for the random variable to be jointly Gaussian, a necessary and sufficient condition is that

- Z_1 and Z_2 are each marginally Gaussian, and
- $Z_1|Z_2 = z$ is Gaussian, and $Z_2|Z_1 = z$ is Gaussian.

A second characterization of a jointly Gaussian zero mean RV $Z \in \mathbb{R}^2$ is that it can be written as $Z = AX$, where $X \in \mathbb{R}^2$ is a collection of i.i.d. standard normal RVs and $A \in \mathbb{R}^{2\times 2}$ is a matrix.

Note that the probability density function of a non-degenerate (meaning the covariance matrix is positive definite and thus invertible) multivariate Gaussian RV with mean vector, μ, and covariance matrix, Σ, is:

$$f(\mathbf{z}) = \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right) / \sqrt{(2\pi)^k|\Sigma|}$$

.

Let X_1 and X_2 be i.i.d. standard normal RVs. Let U denote a binary random variable uniformly distributed on $\{-1, 1\}$, independent of everything else. Use one of the two characterizations given above to determine whether the following RVs are jointly Gaussian, and calculate the covariance matrix (regardless of whether the RVs are jointly Gaussian).

- $Z_1 = X_1$ and $Z_2 = X_2$.
- $Z_1 = X_1$ and $Z_2 = X_1 + X_2$.
- $Z_1 = X_1$ and $Z_2 = -X_1$.
- $Z_1 = X_1$ and $Z_2 = UX_1$.

(b) (2 points) Show that two Gaussian random variables can be uncorrelated, but not independent (*Hint: use one of the examples in part (a)*). On the other hand, show that two uncorrelated, jointly Gaussian RVs are independent.

(c) (1 point) With the setup in (a), let $Z = VX$, where $V \in \mathbb{R}^{2\times 2}$, and $Z, X \in \mathbb{R}^2$. What is the covariance matrix Σ_Z? Is this necessarily true if X has the identity matrix $I \in \mathbb{R}^{2\times 2}$ as its covariance matrix, but is not a multivariate Gaussian?

(d) (2 points) Use the above setup to show that $X_1 + X_2$ and $X_1 - X_2$ are independent. Give another example pair of linear combinations are independent (You may not simply use multiples of $X_1 + X_2$ and $X_1 - X_2$, or X_1 and X_2).

(e) (3 points) Given a jointly Gaussian zero mean RV $Z = [Z_1\ Z_2]^\top \in \mathbb{R}^2$ with covariance matrix $\Sigma_Z = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}$, derive the distribution of $Z_1|Z_2 = z$.

Hint: The following identity may be useful

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{b}{c} & 1 \end{bmatrix} \begin{bmatrix} \left(a - \frac{b^2}{c}\right)^{-1} & 0 \\ 0 & \frac{1}{c} \end{bmatrix} \begin{bmatrix} 1 & -\frac{b}{c} \\ 0 & 1 \end{bmatrix}.$$

2 Linear Regression, Projections and Pseudoinverses (13 points)

We are given $X \in \mathbb{R}^{n \times d}$ where $n > d$ and $\text{rank}(X) = d$. We are also given a vector $y \in \mathbb{R}^n$. Define the orthogonal projection of y onto $\text{range}(X)$ as $P_X(y)$.

Background on orthogonal projections For any finite-dimensional subspace W (here, $\text{range}(X)$) of a vector space V (here, $\mathbb{R}^n$), any vector $v \in V$ can be decomposed as

$$v = w + u, \qquad w \in W, \quad u \in W^\perp,$$

where $W^\perp$ is the orthogonal complement of W. Furthermore, this decomposition is unique: if $v = w' + u'$ where $w' \in W$, $u' \in W^\perp$, then $w' = w$ and $u' = u$. These two facts allow us to define P_W, the orthogonal projection operator onto W. Given a vector v with decomposition $v = w + u$, we define

$$P_W(v) = w.$$

It can also be shown using these two facts that P_W is linear. For more information on orthogonal projections, see `https://gwthomas.github.io/docs/math4ml.pdf`.

(a) (2 points) Prove that $P_X(y) = \underset{w \in \text{range}(X)}{\arg\min} \|y - w\|_2^2$.

Side Note: In lecture, we sketched the geometric intuition of a projection and the connection between minimizing the least squares loss $L(\theta) = \|y - X\theta\|_2^2$ and finding a projection. Least squares seeks the vector in the columnspace of X that is the closest to y. Hence, $P_X(y) = X\theta^*$.

(b) (3 points) An orthogonal projection is a linear transformation. Hence, we can define $P_X(y) = Py$ for some projection matrix P. Specifically, given $1 \leq d \leq n$, a matrix $P \in \mathbb{R}^{n \times n}$ is said to be a rank-d orthogonal projection matrix if $\text{rank}(P) = d$, $P = P^\top$ and $P^2 = P$. Prove that P is a rank-d projection matrix if and only if there exists a $U \in \mathbb{R}^{n \times d}$ such that $P = UU^\top$ and $U^\top U = I$.

Hint Use the eigendecomposition of P.

(c) (1 point) Prove that if P is a rank d projection matrix, then $\text{Tr}(P) = d$.

(d) (2 points) The Singular Value Decomposition theorem states that we can write any matrix X as

$$X = \sum_{i=1}^{\min\{n,d\}} \sigma_i u_i v_i^\top = \sum_{i:\sigma_i > 0} \sigma_i u_i v_i^\top$$

where $\sigma_i \geq 0$, and $\{u_i\}_{i=1}^n$ and $\{v_i\}_{i=1}^d$ are orthonormal bases for $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively. Some of the singular values σ_i may equal 0, indicating that the associated left and right singular vectors u_i and v_i do not contribute to the sum, but sometimes it is still convenient to include them in the SVD so we have complete orthonormal bases for $\mathbb{R}^n$ and $\mathbb{R}^d$ to work with. Show that

(i) $\{v_i : \sigma_i > 0\}$ is an orthonormal basis for the row space of of X

(ii) Similarly, $\{u_i : \sigma_i > 0\}$ is an orthonormal basis for the columnspace of X
Hint: consider $X^\top$.

(e) (2 points) Prove that if $X \in \mathbb{R}^{n \times d}$ and $\text{rank}(X) = d$, then $X(X^\top X)^{-1}X^\top$ is a rank-d orthogonal projection matrix. What is the corresponding matrix U, in terms of an SVD of X, such that $X(X^\top X)^{-1}X^\top = UU^\top$ and $U^\top U = I$?

(f) (3 points) Define the Moore-Penrose pseudoinverse to be the matrix:

$$X^\dagger = \sum_{i:\sigma_i > 0} \sigma_i^{-1} v_i u_i^\top,$$

(i) Show that $X^\dagger X$ is a projection matrix.

(ii) Which subspace does it project onto?

(iii) What is this subspace if $\text{rank}(X) = d$?

(iv) If $\text{rank}(X) = d$ and $n = d$, does X^{-1} exist? If so, how is $X^\dagger$ related to X^{-1}?

3 Some MLEs (11 points)

For this question, assume you observe n (data point, label) pairs $(x_i, y_i)_{i=1}^n$, with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for all $i = 1, \ldots, n$. We denote X as the data matrix containing all the data points and y as the label vector containing all the labels:

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n.$$

(a) (4 points) Ignoring y for now, suppose we model the data points as coming from a d-dimensional Gaussian with diagonal covariance:

$$\forall i = 1, \ldots, n, \quad x_i \overset{i.i.d.}{\sim} N(\mu, \Sigma); \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_d^2 \end{bmatrix}.$$

If we consider $\mu \in \mathbb{R}^d$ and $(\sigma_1^2, \ldots, \sigma_d^2)$, where each $\sigma_i^2 > 0$, to be unknown, the parameter space here is $2d$-dimensional. When we refer to Σ as a parameter, we are referring to the d-tuple $(\sigma_1^2, \ldots, \sigma_d^2)$, but inside a linear algebraic expression, Σ denotes the diagonal matrix $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. Compute the log-likelihood $\ell(\mu, \Sigma) = \log p(X \mid \mu, \Sigma)$, and,

(i) find the MLE of μ assuming Σ is known;

(ii) find the MLE of Σ assuming μ is known;

(iii) find the joint MLE of (μ, Σ).

Justify why your answers maximize the likelihood.

(b) (3 points) Now we consider X as fixed, and not a random matrix. Suppose we model the outcome y as

$$y_i \overset{ind.}{\sim} N(x_i^\top w, \sigma^2) \quad \forall i = 1, \ldots, n,$$

where $w \in \mathbb{R}^d$ is an unknown parameter.

(i) Compute the log-likelihood $\ell(w) = \log p(y \mid w)$ and show that finding the MLE of w is equivalent to linear regression:

$$\min_w \|Xw - y\|_2^2.$$

In lecture it was shown that the unique solution when $\mathrm{rank}(X) = d$ is $w^* = (X^T X)^{-1} X^\top y$. In the case where $\mathrm{rank}(X) < d$, show

(ii) if a solution exists, it is not unique;

(iii) a solution does exist, namely $X^\dagger y$ is a solution.

Hint: the concepts in Question 2 of this homework may help here.

One interesting fact to note is that $X^\dagger y$ is the minimum norm solution. You have the tools you need to prove this, and feel free to optionally include a proof in your writeup, but this will not be graded.

(c) (2 points) Again considering X fixed, now suppose that each $y_i \in \{0, 1\}$ is binary-valued and that we model y as

$$y_i \overset{ind.}{\sim} \text{Ber}(s(x_i^\top w)) \quad \forall i = 1, \ldots, n,$$

where $s : \mathbb{R} \to \mathbb{R}$, $s(z) = \frac{1}{1+e^{-z}}$ is the *sigmoid* function, and $\text{Ber}(p)$ denotes the Bernoulli distribution which takes value 1 with probability p and 0 with probability $1 - p$.

(i) Write down the log-likelihood $\ell(w) = \log p(y \mid w)$ and show that finding the MLE of w is equivalent to minimizing the cross entropy between $\text{Ber}(y_i)$ and $\text{Ber}(s(x_i^\top w))$ for each i:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} H(\text{Ber}(y_i), \text{Ber}(s(x_i^\top w))). \tag{1}$$

Definition of cross entropy: given two discrete probability distributions $\pi : \Omega \to [0, 1]$ and $\theta : \Omega \to [0, 1]$ on some outcome space Ω, we have $\sum_{\omega \in \Omega} \pi(\omega) = \sum_{\omega \in \Omega} \theta(\omega) = 1$ and

$$H(\pi, \theta) = \sum_{\omega \in \Omega} -\pi(\omega) \log \theta(\omega).$$

(ii) Show that (1) (and therefore finding the MLE) is equivalent to the following problem:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \log\Big(1 + \exp\big(-z_i x_i^\top w\big)\Big) \tag{2}$$

where $z_i = 1$ if $y_i = 1$ and $z_i = -1$ if $y_i = 0$.
Note: both (1) and (2) are referred to as logistic regression.

(d) (2 points) Consider the Categorical$(\theta_1, \ldots, \theta_K)$ distribution ($\theta_k \geq 0 \, \forall k$ and $\sum_{k=1}^K \theta_k = 1$), which takes values in $\{1, \ldots, K\}$. A random variable Z with this distribution has

$$P(Z = k) = \theta_k \quad \forall k = 1, \ldots, K.$$

Now, ignoring the data points X, suppose that $\forall i = 1, \ldots, n, \; y_i \in \{1, \ldots, K\}$, and

$$y_i \overset{i.i.d.}{\sim} \text{Categorical}(\theta_1, \ldots, \theta_K).$$

Compute the MLE of $\theta = (\theta_1, \ldots, \theta_K)$. One method is to use Lagrange multipliers. Another method is to use the fact that the KL divergence is nonnegative:

$$\text{KL}(\pi \,\|\, \theta) = \sum_{\omega \in \Omega} \pi(\omega) \log\left(\frac{\pi(\omega)}{\theta(\omega)}\right) \geq 0.$$

You are free to use either method to compute the MLE of θ.

4 Geometry of Ridge Regression (13 points)

You recently learned ridge regression and how it differs from ordinary least squares. In this question we will explore how ridge regression is related to solving a constrained least squares problem in terms of their parameters and solutions.

(a) (1 point) Given a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and a vector $\mathbf{y} \in \mathbb{R}^n$, define the optimization problem

$$\begin{aligned} &\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \ \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2. \\ &\text{subject to } \|\mathbf{w}\|_2^2 \leq \beta^2. \end{aligned} \tag{3}$$

We can utilize Lagrange multipliers to incorporate the constraint into the objective function by adding a term which acts to "penalize" the thing we are constraining. Write down the *Lagrangian* of this constrained problem.

Relevant background on Lagrangians: Given a constrained problem of the form

$$\begin{aligned} &\underset{\mathbf{x} \in \mathbb{R}^d}{\textit{minimize}} \quad f_0(\boldsymbol{x}) \\ &\textit{subject to} \quad f_i(\boldsymbol{x}) \leq c_i \quad \forall\, i = 1, \ldots, p, \end{aligned}$$

where $f_0, \ldots, f_p$ are functions $\mathbb{R}^d \to \mathbb{R}$ and $c_1, \ldots, c_p$ are scalars, its Lagrangian is defined as

$$\mathcal{L}(\boldsymbol{x}, \lambda) = f_0(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i (f_i(\boldsymbol{x}) - c_i).$$

According to duality theory, if all the functions $f_0, \ldots, f_p$ are convex and there exists a strictly feasible point (this happens to be true for our problem, assuming $\beta > 0$), then there exists $\lambda^* \in \mathbb{R}^p$ such that

$$\arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \lambda^*)$$

is the solution to the original constrained problem.

(b) (1 point) Recall that ridge regression is given by the unconstrained optimization problem

$$\min_{w} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \nu \|\mathbf{w}\|_2^2. \tag{4}$$

This means that one way to interpret "ridge regression" is as the Lagrangian form of a constrained problem. Qualitatively, how would increasing β in our previous problem be reflected in the desired penalty ν of ridge regression (i.e. if our threshold β increases, what should we do to ν)?

(c) (1 point) One reason why we might want to have small weights $\mathbf{w}$ has to do with the sensitivity of the predictor to its input. Let $\mathbf{x}$ be a d-dimensional list of features corresponding to a new test point. Our predictor is $\mathbf{w}^\top \mathbf{x}$. What is an upper bound on how much our prediction could change if we added noise $\boldsymbol{\epsilon} \in \mathbb{R}^d$ to a test point's features $\mathbf{x}$, in terms of $\|\mathbf{w}\|_2$ and $\|\boldsymbol{\epsilon}\|_2$?

(d) (2 points) Derive that the solution to ridge regression (4) is given by $\hat{\mathbf{w}}_\mathbf{r} = (\mathbf{X}^\top\mathbf{X} + \nu\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$. What happens when $\nu \to \infty$? It is for this reason that sometimes regularization is referred to as "shrinkage."

(e) (1 point) Note that in computing $\hat{\mathbf{w}}_\mathbf{r}$, we are trying to invert the matrix $\mathbf{X}^\top\mathbf{X} + \nu\mathbf{I}$ instead of the matrix $\mathbf{X}^\top\mathbf{X}$. If $\mathbf{X}^\top\mathbf{X}$ has eigenvalues $\sigma_1^2, \ldots, \sigma_d^2$, what are the eigenvalues of $\mathbf{X}^\top\mathbf{X} + \nu\mathbf{I}$? Comment on why adding the regularizer term $\nu\mathbf{I}$ can improve the inversion operation numerically.

(f) (1 point) Let the number of parameters $d = 3$ and the number of datapoints $n = 5$, and let the eigenvalues of $\mathbf{X}^\top\mathbf{X}$ be given by 1000, 1 and 0.001. We must now choose between two regularization parameters $\nu_1 = 100$ and $\nu_2 = 0.5$. Which do you think is a better choice for this problem and why?

(g) (2 points) Another advantage of ridge regression can be seen for under-determined systems. Say we have the data drawn from a $d = 5$ parameter model, but only have $n = 4$ training samples of it, i.e. $\mathbf{X} \in \mathbb{R}^{4\times5}$. Now this is clearly an underdetermined system, since $n < d$. Show that ridge regression with $\nu > 0$ results in a unique solution, whereas ordinary least squares has an infinite number of solutions.

Hint: To make this point, it may be helpful to consider $\mathbf{w} = \mathbf{w}_0 + \mathbf{w}^*$ where $\mathbf{w}_0$ is in the null space of $\mathbf{X}$ and $\mathbf{w}^*$ is a solution.

(h) (2 points) What will the solution to ridge regression given in part (d) converge to if you take the limit $\nu \to 0$?

Hint: Use the SVD of $\mathbf{X}$.

(i) (2 points) Tikhonov regularization is a general term for ridge regression, where the implicit constraint set takes the form of an ellipsoid instead of a ball. In other words, we solve the optimization problem

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \nu\|\Gamma\mathbf{w}\|_2^2$$

for some full rank matrix $\Gamma \in \mathbb{R}^{d\times d}$. Derive a closed form solution for $\mathbf{w}$.

5 Robotic Learning of Controls from Demonstrations and Images (11 points)

Huey, a home robot, is learning to retrieve objects from a cupboard, as shown in Fig. 1. The goal is to push obstacle objects out of the way to expose a goal object. Huey's robot trainer, Anne, provides demonstrations via tele-operation. When tele-operating the robot, Anne can look at the images captured by the robot and provide controls to Huey remotely.

During a demonstration, Huey records the RGB images of the scene for each of the n timesteps, $x_1, ..., x_n$, where $x_i \in \mathbb{R}^{30\times30\times3}$ and the controls for his body for each of the n timesteps, $u_1, \ldots, u_n$, where $u_i \in \mathbb{R}^3$. The controls correspond to making small changes in the 3D pose (i.e. translation and rotation) of his body. Examples of the data are shown in the figure.

Under an assumption (sometimes called the Markovian assumption) that all that matters for the current control is the current image, Huey can try to learn a linear *policy* π (where $\pi \in \mathbb{R}^{2700\times3}$) which linearly maps image states to controls (i.e. $\pi^\top x = u$). We will now explore how Huey can recover this policy using linear regression. Note please use **numpy** and **numpy.linalg** to complete this assignment.

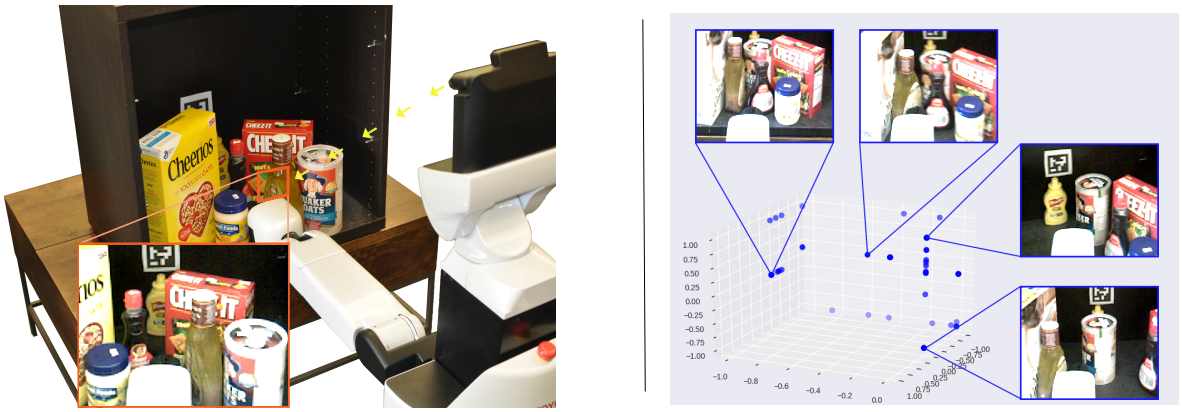

Figure 1: A) Huey trying to retrieve a mustard bottle. An example RGB image of the workspace taken from his head mounted camera is shown in the orange box. The angle of the view gives Huey and eye-in-hand perspective of the cupboard he is reaching into. B) A scatter plot of the 3D control vectors, or u labels. Notice that each coordinate of the label lies within the range of $[-1, 1]$ for the change in position. Example images, or states x, are shown for some of the corresponding control points. The correspondence is indicated by the blue lines.

(a) (2 points) To get familiar with the structure of the data, **please visualize the 0th, 10th and 20th images in the training dataset. Also find out what's their corresponding control vectors.**

(b) (1 points) Load the n training examples from x_train.p and compose the matrix X, where $X \in \mathbb{R}^{n\times2700}$. Note, you will need to flatten the images to reduce them to a single vector. The flattened image vector will be denoted by $\bar{x}$ (where $\bar{x} \in \mathbb{R}^{2700\times1}$). Next, load the n examples

from y_train.p and compose the matrix U, where $U \in \mathbb{R}^{n\times 3}$. Try to perform ordinary least squares by forming the matrix $(X^\top X)^{-1}X^\top$ to solve

$$\min_{\pi} ||X\pi - U||_F$$

to learn the *policy* $\pi \in \mathbb{R}^{2700\times 3}$. **Report what happens as you attempt to do this and explain why.**

(c) (2 points) Now try to perform ridge regression:

$$\min_{\pi} ||X\pi - U||_F^2 + \lambda||\pi||_F^2$$

on the dataset for regularization values $\lambda = \{0.1, 1.0, 10, 100, 1000\}$. Measure the average squared Euclidean distance for the accuracy of the policy on the training data:

$$\frac{1}{n}\sum_{i=0}^{n-1} ||\bar{x}_i^T\pi - u_i^\top||_2^2$$

(In the above, we are taking the ℓ_2 norm of a row vector, which here we take to mean the ℓ_2 norm of the column vector we get by transposing it.) **Report the training error results for each value of λ.**

(d) (2 points) Next, we are going to try standardizing the states. For each pixel value in each data point, x, perform the following operation:

$$x \mapsto \frac{x}{255}\times 2 - 1.$$

Since we know the maximum pixel value is 255, this rescales the data to be between $[-1, 1]$. **Repeat the previous part and report the average squared training error for each value of λ.**

(e) (3 points) Evaluate both *policies* (i.e. with and without standardization on the new validation data x_test.p and y_test.p for the different values of λ. **Report the average squared Euclidean loss** and **qualitatively explain how changing the values of λ affects the performance in terms of bias and variance**.

(f) (1 point) To better understand how standardizing improved the loss function, we are going to evaluate the *condition number* κ of the optimization, which is defined as

$$\kappa = \frac{\sigma_{\max}(X^TX + \lambda I)}{\sigma_{\min}(X^TX + \lambda I)}$$

or the ratio of the maximum singular value to the minimum singular value of the relevant matrix. Roughly speaking, the condition number of the optimization process measures how stable the solution will be when some error exists in the observations. More precisely, given

a linear system $Ax = b$, the condition number of the matrix A is the maximum ratio of the relative error in the solution x to the relative error of b.

For the regularization value of $\lambda = 100$, **report the condition number with the standardization technique applied and without**.