CS/ECE/ME 532
Extra Credit Problems

1. **Gradient Descent and Stochastic Gradient Descent**. Suppose we have training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^m$, with $\boldsymbol{x}_i \in \mathbb{R}^n$ and y_i is a scalar label. Derive gradient descent and SGD algorithms to solve the following ℓ_1-loss optimization:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \sum_{i=1}^m \left| y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{w} \right| .$$

   a) Simulate this problem as follows. Generate each x_i as random points in the interval $[0, 1]$ and generate $y_i = w_1 x_i + w_2 + \epsilon_i$, where w_1 and w_2 are the slope and intercept of a line (of your choice) and $\epsilon_i =$ `randn`, a Gaussian random error generated in Matlab. With $m = 10$. Repeat this experiment with several different datasets (with different random errors in each case).

   b) Implement the GD or SGD algorithm for the ℓ_1-loss optimization. Compare the solution to this optimization with the LS line fit.

   c) Now change the simulation as follows. Instead of generating ϵ_i as Gaussian, now generate the errors according to a Laplacian (two-sided exponential distribution) using `laprnd(1,1)`. Compare the LS and ℓ_1-loss solution compare in this case. Repeat this experiment with several different datasets (with different random errors in each case).

2. **Classification and the SVM**. Revisit the `iris` data set from Homework 3. For this problem, we will use the 3rd and 4th features to classify whether an iris is *versicolor* or *virginica*. Here is a plot of the data set for this restricted set of features.

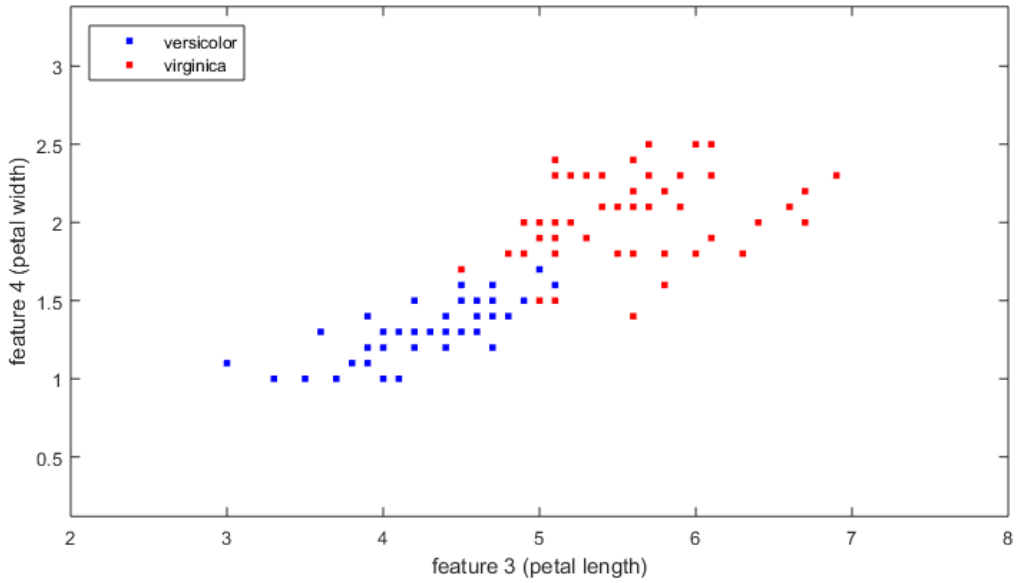
   We will look for a linear classifier of the form: $x_{i3}w_1 + x_{i4}w_2 + w_3 \approx y_i$. Here, x_{ij} is the measurement of the j^{th} feature of the i^{th} iris, and w_1, w_2, w_3 are the weights we would like to find. The y_i are the labels; e.g. $+1$ for *versicolor* and -1 for *virginica*.

   a) Reproduce the plot above, and also plot the decision boundary for the least squares classifier.

   b) This time, we will use a regularized SVM classifier with the following loss function:

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \sum_{i=1}^m (1 - y_i \boldsymbol{x}_i^\mathsf{T}\boldsymbol{w})_+ + \lambda(w_1^2 + w_2^2)$$

   Here, we are using the standard hinge loss, but with an ℓ_2 regularization that penalizes only w_1 and w_2 (we do not penalize the offset term w_3). Solve the problem by implementing gradient descent of the form $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \gamma\nabla f(\boldsymbol{w}_t)$. For your numerical simulation, use parameters $\lambda = 0.1$, $\gamma = 0.003$, $\boldsymbol{w}_0 = \boldsymbol{0}$ and $T = 20{,}000$ iterations. Plot the decision boundary for this SVM classifier. How does it compare to the least squares classifier?



c) Let's take a closer look at the convergence properties of $\boldsymbol{w}_t$. Plot the three components of $\boldsymbol{w}_t$ on the same axes, as a function of the iteration number t. Do the three curves each appear to be converging? Now produce the same plots with a larger stepsize ($\gamma = 0.01$) and a smaller stepsize ($\gamma = 0.0001$). What do you observe?

3. Projection matrices. A square matrix P is called a *projection matrix* if $P^2 = P$.

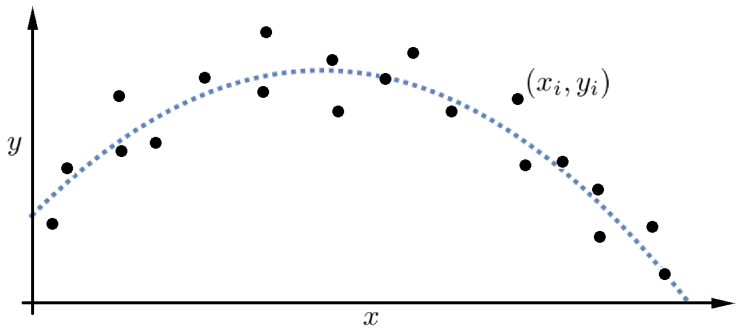
a) Prove that if P is a projection matrix, so is $I - P$, where I is the $n \times n$ identity matrix.

b) Prove that if A has linearly independent columns, then $A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ is a projection matrix.

c) Prove that if $\{u_1, \ldots, u_k\}$ are orthonormal vectors, then $u_1u_1^\mathsf{T} + \cdots + u_ku_k^\mathsf{T}$ is a projection matrix.

4. Baseball pitching machine. You've just built your first prototype for a baseball pitching machine and now you'd like to model the trajectory of the baseballs launched from the machine. You have a GPS tracking beacon at your disposal so you attach it to a baseball, launch it, collect data, and repeat. Unfortunately, the GPS device is crude: it records distances and altitudes (x_i, y_i), $i = 1, 2, \ldots, n$, but the data are noisy and you only get a few samples per launch. The data points after several launches are shown in the figure below.

a) You expect the true baseball trajectory to resemble a parabola with equation $y = px^2 + qx + r$ for some choice of p, q, and r. An example of such a curve is shown as a dotted line in the figure. Describe (1) how you would formulate this as a least-squares problem to find p, q, and r that best fit the data you've gathered and (2) how you would



solve this least squares problem. **Note:** This is not an idealized problem! There are aerodynamic effects in play so don't assume anything about physics (e.g. knowledge of Earth's gravity). All you have is the data!

(continued on next page)

b) Suppose you know your baseballs are being launched from an altitude of $y = 1$ (when $x = 0$). How should you modify your least-squares problem to account for this?

c) In addition, your machine is oriented such that the baseballs are launched at an angle of exactly 45 degrees. How should you modify your least-squares problem to account for this?

5. Visualizing least squares. Consider the problem of finding x that minimizes $\|Ax - b\|$ where:

$$A = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

a) What is the solution $\hat{x}$ to the least squares problem above? Also compute the corresponding optimal projection $\hat{b} := A\hat{x}$ and optimal residual $\hat{r} := b - A\hat{x}$.

b) Make a plot of $\mathbb{R}^2$ showing the subspaces range(A) and range$(A)^\perp$. On your plot, also include and label the vectors b, $\hat{b}$, and $\hat{r}$.

6. A microbiology experiment. We have cultured a new bacterial strain and we would like to predict growth conditions under which the bacteria express a certain gene. We perform several experiments where we vary the abundance of two key nutrients and then measure gene expression. The results:

Experiment number (i)	Relative abundance of nutrient 1 (p_i)	Relative abundance of nutrient 2 (q_i)	Expression of the target gene (y_i)
1	+0.5	+0.5	+1
2	−0.5	0	−1
3	0	−0.5	−1

The nutrient abundance is normalized, so a value of 0 is average, positive values indicate increased levels, and negative values indicate decreased values. For the gene expression, $+1$ means the gene was expressed and -1 means the gene was not expressed.

a) Find weights w_1 and w_2 such that $y_i = \text{sign}(p_i w_1 + q_i w_2)$ for every experiment i. Plot the points $(p_i, q_i) \in \mathbb{R}^2$ along with the decision boundary determined by your chosen weights.
Hint: you should be able to do this by inspection; no calculations are needed!

b) We perform a fourth experiment in which $p_4 = -0.1$, $q_4 = -0.1$, and $y_4 = +1$. Explain why it's no longer possible to find w_1 and w_2 that achieve perfect classification as before. Suggest a simple modification to the classification rule that allows for perfect classification.

7. Properties of the SVD. Here is a 3×2 matrix A along with its singular value decomposition:

$$A = \begin{bmatrix} 8 & 19 \\ -2 & 14 \\ 20 & 10 \end{bmatrix} = U\Sigma V^\mathsf{T}, \quad \text{with:} \quad U = \frac{1}{3}\begin{bmatrix} 2 & 1 & -2 \\ 1 & 2 & 2 \\ 2 & -2 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 30 & 0 \\ 0 & 15 \\ 0 & 0 \end{bmatrix}, \quad V = \frac{1}{5}\begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix}.$$

All parts of this problem refer to the specific matrix A above.

a) List four properties that make this a legitimate SVD. You do not need to verify all of these properties; just write down what the properties are.

b) The SVD is not unique. Write down an SVD for A that is different from the one above. By *different*, I mean that at least one of the matrices U, Σ, V is different from the ones above.

c) Find a nonzero vector $x \in \mathbb{R}^2$ such that when you compute the product $y = Ax$, the amplification factor $\|y\|/\|x\|$ is **as small as possible**. Also state the corresponding amplification factor.
Properties of the SVD (cont'd). These problems refer to the same A as in the prev. page.

d) What is the rank 1 matrix that approximates A in the sense of minimizing the Frobenius norm? In other words, find the $X \in \mathbb{R}^{3\times 2}$ that solves

$$\begin{aligned} \text{minimize} \quad & \|X - A\|_F \\ \text{subject to:} \quad & (X) = 1 \end{aligned}$$

e) Consider the least squares problem of finding $x \in \mathbb{R}^2$ that minimizes $\|Ax - b\|$ where A is the matrix given in the start of this problem. Describe the set of all vectors b such that the least squares solution has a zero residual (in other words, $A\hat{x} = b$).

f) Again consider the least squares problem from the previous part. Describe the set of all vectors b such that the optimal least squares solution is $\hat{x} = 0$.

Alternatively, A is full column rank so saying that $\hat{x} = 0$ is equivalent to saying that $A\hat{x} = 0$. So in other words, the best approximation $\hat{b} \in \text{range}(A)$ to the vector b is zero. This implies that $b \in \text{range}(A)^{\perp}$, and so b must lie in the subspace determined by the last column of U.

8. Trade-offs. We are trying to find an $x \in \mathbb{R}$ that is (i) close to 1 and (ii) not too big. We formulate this as a one-dimensional L_2-regularized least squares problem (with $\lambda \geq 0$)

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad (x-1)^2 + \lambda x^2$$

a) What is the optimal $\hat{x}$ (as a function of λ)?

b) For the optimal $\hat{x}$, carefully plot the trade-off curve with $\hat{x}^2$ on the x-axis and $(\hat{x}-1)^2$ on the y-axis. ***Hint:*** chose a few λ values and connect the dots.

c) Suppose we use the same solutions $\hat{x}$ as above but this time we plot the trade-off curve with $|\hat{x}|$ on the x-axis and $|\hat{x}-1|$ on the y-axis. What is the shape of this different trade-off curve, and why does it have this shape?