6.7900 Fall 2024: Homework 0

This is a set of diagnostic and warm-up problems, divided into two parts.

1. *To be handed in*: Problems in the first part represent basic background in linear algebra, applied math, and optimization. They aren't trivial, but if they aren't relatively easy for you, then it might be better to gain some more background in a prerequisite area before taking 6.7900.

2. *To be used for your own practice*: Problems in the second part are designed to help you learn/practiced numpy-style "vectorized" programming strategies, which will help you create efficient implementations of algorithms studied in class and also to interface with existing libraries. Please solve these problems using numpy, striving for elegant and efficient solutions.

   If you don't have a Python/numpy installation on your own computer (or even if you do!) Google Colab (`https://colab.google/`) is a good way to get going quickly.

Submission instructions

Please hand in your work via Gradescope via the link at `https://gradml.mit.edu/info/homeworks/`.

- Latex is not required, but if you are hand-writing your solutions, **please** write clearly and carefully. You should include enough work to show how you derived your answers, but you don't have to give careful proofs.
- Homework is due on Tuesday September 10 at 11PM.
- Lateness and extension policies are described at `https://gradml.mit.edu/info/class_policy/`.

1 Math Background

1.1 Just plane fun

Consider a hyperplane in n-dimensional Euclidean space, described by the $n+1$ real values w_i for $i = 0, \ldots, n$: the hyperplane consists of points $(x_1, \ldots, x_n)$ satisfying

$$w_0 + w_1 x_1 + \ldots + w_n x_n = 0 \ .$$

1. Find a unit vector normal to the hyperplane. Given a point $\boldsymbol{v} = (v_1, \ldots, v_n)$ on the hyperplane, give the equation for the line through the point that is orthogonal to the hyperplane.
2. Given a point $\boldsymbol{v} = (v_1, \ldots, v_n)$, how can you determine which side of the hyperplane it is on?
3. What is the distance of a point $\boldsymbol{v} = (v_1, \ldots, v_n)$ to the hyperplane?

1.2 Multivariate Gaussian

Let X be a random variable taking values in $\mathbb{R}^n$. It is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Recall that the probability density function (pdf) $p_X(\boldsymbol{x})$, sometimes denoted $p(X = \boldsymbol{x})$, for X is given by

$$p_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

1. Show how to obtain the normalization constant $1/\sqrt{(2\pi)^n |\Sigma|}$ for the multivariate Gaussian, starting from the fact that

   $$p_X(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

   Hints: It's fine to assume $\boldsymbol{\mu} = 0$ (Why?). A useful (and cool!) fact is that $\int_{\mathbb{R}} \exp(-\frac{1}{2}x^2) = \sqrt{2\pi}$.
2. Let the random variable $Y = 2X$. What is the pdf of Y?
3. What can we say about the distribution of X if Σ is the identity matrix, I? Does this imply anything about factorization of the pdf?
4. What can we say about the distribution of X if Σ is $\begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$? Approximately what shape do equiprobability contours (i.e., sets $\{x \in R^n : p_X(x) = c\}$ for some c) have?
5. What can we say about the distribution of X if Σ is $\begin{bmatrix} 10 & -4 \\ -4 & 10 \end{bmatrix}$? Approximately what shape do equiprobability contours of this distribution have?

6. Is $\begin{bmatrix} 2 & 10 \\ 10 & 2 \end{bmatrix}$ a valid Σ? How can you tell?

7. If $\mu = (1, 2)$, and $\Sigma = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$, what is the conditional pdf $p_{X_1|X_2}(x_1 \mid 3)$?

1.3 Probability

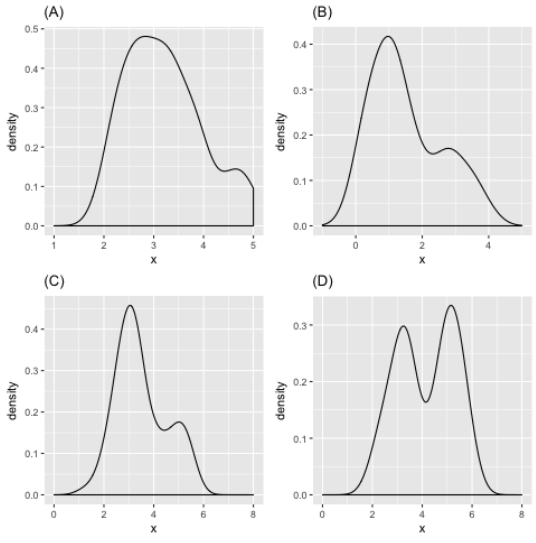
1. Let A, B be $p \times q$ matrices and x be a random $q \times 1$ vector. Prove that

$$\text{cov}(Ax, Bx) = A\,\text{cov}(x)B^T$$

where $\text{cov}(u, v) = E\left[(u - E[u])(v - E[v])^T\right]$ is the cross-covariance matrix between random vectors u and v, while $\text{cov}(u) = E\left[(u - E[u])(u - E[u])^T\right]$ is the covariance matrix for u.

2. You go for your annual checkup and have several lab tests performed. A week later your doctor calls you and says she has good and bad news. The bad news is that you tested positive for a marker of a serious disease, and that the test is 97% accurate (i.e. the probability of testing positive given that you have the disease is 0.97, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only 1 in 20,000 people. What are the chances that you actually have the disease?

3. Consider the following generative process describing the joint distribution $p(Z, X)$: $Z \sim$ Bernoulli$(0.2), X \sim$ Gaussian$(\mu_Z, 0.5)$, where $\mu_0 = 3$ and $\mu_1 = 5$. Which of the following plots is the marginal distribution $p(X)$?



4. Alice and Bob were driving through a tunnel while listening to the Billion-dollar lottery drawing live on the radio. Due to the weak signal, they couldn't hear the last number perfectly clearly. Alice and Bob think the number was A and B, respectively. Is A independent of B? Is A independent of B given the true lottery number T?

1.4 Multivariate calculus

1. Find the minimum value of the function $f(x, y) = x^2 + 2y^2 - xy + x - 4y$ over $\mathbb{R}^2$.
2. Show that $f(x, y)$ is convex over $\mathbb{R}^2$ by showing that its Hessian is positive semi-definite.

Note: Here's a handy tool for quickly (sanity) check your matrix calculus `http://matrixcalculus.org/`.

1.5 Optimization and Gradient Descent

Grady Ent decides to train a single sigmoid unit using the following objective function:

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_i (\sigma(\boldsymbol{x}^{(i)} \cdot \boldsymbol{w}) - y^{(i)})^2 + \frac{\beta}{2}\sum_j w_j^2$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function. Note that $\boldsymbol{x}^{(i)} \cdot \boldsymbol{w}$ is the inner product between the vectors $\boldsymbol{x}^{(i)}$ and $\boldsymbol{w}$, where $(\boldsymbol{x}^{(i)}, y^{(i)})$ is the i-th training data point.

(a) Write an expression for $\partial E/\partial w_j$.

(b) Give the gradient descent update to weight w_j given a single training example $(\boldsymbol{x}, y)$. Your answer should be in terms of the training data and a learning rate.

(c) Is the following claim true? "Stochastic gradient descent steps always decrease the objective". Please provide a brief justification for your answer.

(d) Is the following claim true? "There are circumstances in which stochastic gradient descent is to be preferred to exact gradient descent". Please provide a brief justification for your answer.

2 Programming problems: Just for practice — do not turn in!

Do not use `for` loops in any of these solutions! (Some of these are tricky, but cool when you see it!).

2.1 Regularization

Given an $n \times n$ matrix C, add a scalar a to each diagonal entry of C.

2.2 Largest Off-diagonal Element

Given an $n \times n$ matrix A, find the value of the largest off-diagonal element.

2.3 Pairwise Computation

Given a vector x of length m, and a vector y of length n, compute $m \times n$ matrices: A and B, such that $A(i,j) = x(i) + y(j)$, and $B(i,j) = x(i) \cdot y(j)$.

2.4 Pairwise Euclidean Distances

Given a $d \times m$ matrix X, and a $d \times n$ matrix Y, compute an $m \times n$ matrix D, such that $D(i,j) = \|x^i - y^j\|$, where x^i is the i-th column of X, and y^j is the j-th column of Y. Hint: You may find the following decomposition (of the norm/distance squared) helpful for improving your code efficiency:

$$\left\|x^i - y^j\right\|^2 = \sum_{k=1}^{d} \left(x_k^i - y_k^j\right)^2 = \sum_{k=1}^{d} x_k^{i^2} + \sum_{k=1}^{d} y_k^{j^2} - \sum_{k=1}^{d} 2x_k^i y_k^j.$$

2.5 Compute Mahalanobis Distances

The Mahalanobis distance is a measure of the distance between a point P and a distribution D, introduced by P. C. Mahalanobis in 1936. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. This distance is zero if P is at the mean of D, and grows as P moves away from the mean: Along each principal component axis, it measures the number of standard deviations from P to the mean of D. If each of these axes is rescaled to have unit variance, then Mahalanobis distance corresponds to standard Euclidean distance in the transformed space. Mahalanobis distance is thus unitless and scale-invariant, and takes into account the correlations of the data set (from `http://en.wikipedia.org/wiki/Mahalanobis_distance`). Given a center vector c, a positive-definite covariance matrix S, and a set of n vectors as columns in matrix X, compute the distances of each column in X to c, using the following formula:

$$D(i) = \left(x^i - c\right)^T S^{-1} \left(x^i - c\right). \tag{2}$$

Here, D is a vector of length n.

2.6 2-D Gaussian

Generate 1000 random points from a 2-D Gaussian distribution with mean $\mu = [4, 2]$ and covariance

$$\Sigma = \begin{pmatrix} 1 & 1.5 \\ 1.5 & 3 \end{pmatrix} \tag{3}$$

Plot the points so obtained, and estimate their mean and covariance from the data. Find the eigenvectors of the covariance matrix and plot them centered at the sample mean.

2.7 Tournament fun

A tennis tournament starts with sixteen players. Let's call them $h_i, i = 1, 2, \ldots 16$ (human i, to avoid the potentially confusing notation p_i). The first round has eight games, randomly drawn/paired; i.e., every player has an equal chance of facing any other player. The eight winners enter the next round.

As an enthusiastic tennis and data fan, you have an internal model of these 16 players based on their past performance. In particular, you view each player h_i as having a performance index score $s_i \sim$ Gaussian(θ_i, σ_i^2). The mean θ_i roughly captures the player's 'intrinsic ability' and the variance σ_i^2 roughly captures the player's performance reliability (accounting for recent injuries etc.). In a match between h_i and h_j, player h_i wins if $s_i > s_j$.

Based on your model, what's the probability that your "top seed player" (the one with the highest θ) enters the next round? Run 10,000 simulations to check if it agrees with your answer.