Your Name
email@fas.harvard.edu
CS281-F17

Assignment #0
Due: 5:00pm DATE TBD

Collaborators: John Doe, Fred Doe

Homework 0: Preliminary - Solutions

Grading Instructions

In the solutions, you will see several highlighted checkpoints. These each have a label that corresponds to an entry in the Canvas quiz for this problem set. The highlighted statement should clearly indicate the criteria for being correct on that problem. If you satisfy the criteria for a problem being correct, mark "Yes" on the corresponding position on the Canvas quiz. Otherwise, mark "No". Your homework scores will be verified by course staff at a later date.

That being said, many of the problems in this course will be proofs. If you find a proof that isn't referred to by the course solution, don't worry. If you're uncertain about the proof, make a Piazza post. If you're certain, mark it correct (and we'll look at it during verification).

Introduction

There is a mathematical component and a programming component to this homework. Please submit your PDF and Python files to Canvas, and push all of your work to your GitHub repository. If a question requires you to make any plots, please include those in the writeup.

This assignment is intended to ensure that you have the background required for CS281. You should be able to answer the problems below without complicated calculations. All questions are worth $70/6 = 11.\bar{6}$ points unless stated otherwise.

Variance and Covariance

Problem 1

Let X and Y be two independent random variables.

(a) Show that the independence of X and Y implies that their covariance is zero.

(b) Zero covariance *does not* imply independence between two random variables. Give an example of this.

(c) For a scalar constant a, show the following two properties:

$$\mathbb{E}(X + aY) = \mathbb{E}(X) + a\mathbb{E}(Y)$$
$$\text{var}(X + aY) = \text{var}(X) + a^2\text{var}(Y)$$

Solution

Let $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$.

(a) We note that since μ_X and μ_Y are fixed, then $X \perp Y$ implies that $(X - \mu_X) \perp (Y - \mu_Y)$. Therefore:

$$\begin{aligned}\text{cov}(X, Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \\ &= \mathbb{E}(X - \mu_X)\mathbb{E}(Y - \mu_Y) \\ &= 0\end{aligned}$$

where the second line is by the independence stated above and the final line is because linearity of expectation gives that $E(X - \mu_X) = E(X) - E(\mu_X) = \mu_X - \mu_X = 0$.

Check 1.1: You used the definition of covariance and the fact that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ to conclude that the covariance was 0.

(b) Note that the definition of independence requires the probability distribution to factor, while 0 covariance just requires the expectation to factor, which is much weaker. An example is: take a random variable A as the result of a coin flip, signed ± 1. Then, let B be a random variable that's $1/-1$ if $A = 1$, and equal to 0 otherwise. Then they are clearly dependent (since $p(A = -1, B = -1) = 0$, but their marginal probabilities are not 0), but the covariance is:

$$\mathbb{E}[AB] - \mathbb{E}[A]\mathbb{E}[B] = \left(\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0\right) - 0 = 0$$

Check 1.2: You gave a valid example of two random variables that have 0 covariance, but are not indepedent.

(c) The first statement is linearity of expectation. Let's assume that X, Y are continuous variables:

$$\begin{aligned}\mathbb{E}[X + aY] : &= \int_{(x,y)} (x + ay) p_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_{(x,y)} x p_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y + \int_{(x,y)} ay p_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_x x p_X(x)\, \mathrm{d}x + a \int_y y p_Y(y)\, \mathrm{d}y \\ &= \mathbb{E}[X] + a\mathbb{E}[Y]\end{aligned}$$

Check 1.3: You used either a continuous or discrete definition of the random variables and showed linearity of independence. If you used the continuous form, you used a double integral, and if discrete, you used a double sum, since the original expectation is over the joint. Other solutions are possible.

For the second, we use the fact that for a random variable Z, $\text{var}(Z) = \mathbb{E}(Z^2) - \mathbb{E}(Z)^2$. Setting $Z = X + aY$, we obtain:

$$\begin{aligned}
\text{var}(X + aY) &= \mathbb{E}(X + aY)^2 - (\mathbb{E}(X + aY))^2 \\
&= \mathbb{E}(X^2 + 2aXY + a^2Y^2) - (\mu_X + a\mu_Y)^2 \\
&= (\mathbb{E}(X^2) - \mu_X^2) + a^2(\mathbb{E}(Y^2) - \mu_Y^2) + 2a(\mathbb{E}(XY) - \mu_X\mu_Y) \\
&= \text{var}(X) + a^2\text{var}(Y) + 0
\end{aligned}$$

where the last line uses the independence of X and Y to obtain that $\mathbb{E}(XY) = \mu_X\mu_Y$.

Note: this is a special case of a more general result, which is that for two random variables X, Y that may not be independent, $\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\text{cov}(X, Y)$.

Check 1.4: You expanded the definition of covariance, and used the fact that X, Y are independent.

Densities

Problem 2

Answer the following questions:

(a) Can a probability density function (pdf) ever take values greater than 1?

(b) Let X be a univariate normally distributed random variable with mean 0 and variance $1/100$. What is the pdf of X?

(c) What is the value of this pdf at 0?

(d) What is the probability that $X = 0$?

(e) Explain the discrepancy.

Solution

(a) Yes. Think of densities as "relative" probabilities. Densities must be nonnegative everywhere, and the integral of the PDF over the entire space should equal one.

Check 2.1: You said 'yes', and gave a justification.

(b) The density is $f_X(x) = \frac{10}{\sqrt{2\pi}} e^{-50x^2}$.

Check 2.2: You gave the correct form of the density.

(c) Setting $x = 0$ above yields $f_X(0) = 10/\sqrt{2\pi} \approx 3.99 > 1$.

Check 2.3: You got an answer of $10/\sqrt{2\pi}$.

(d) Since X is a continuous random variable, the probability that it takes on any fixed value is 0.

Check 2.4: You got 0.

(e) The definition of a probability density function is that it is the derivative of the cumulative distribution function. The key property that this implies is that the integral of f_X over a set A equals the probability that $X \in A$. Therefore, the pdf can take on arbitrary values, including values greater than 1, provided its integral over any set is never greater than 1. In particular, the integral of the pdf over the entire support of X equals 1, which can be verified by integrating the density f_X over the entire real line.

Check 2.5: You explained the discrepancy, and were clear that the value of the pdf is not a probability.

Conditioning and Bayes' rule

Problem 3

Let $\mu \in \mathbb{R}^m$ and $\mathbf{\Sigma}, \mathbf{\Sigma}' \in \mathbb{R}^{m \times m}$. Let X be an m-dimensional random vector with $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, and let Y be a m-dimensional random vector such that $Y \mid X \sim \mathcal{N}(X, \mathbf{\Sigma}')$. Derive the distribution and parameters for each of the following.

(a) The unconditional distribution of Y.

(b) The joint distribution for the pair (X, Y).

Hints:

- You may use without proof (but they are good advanced exercises) the closure properties of multivariate normal distributions. Why is it helpful to know when a distribution is normal?
- Review Eve's and Adam's Laws, linearity properties of expectation and variance, and Law of Total Covariance.

Solution

The key to this problem is to know that the closure properties of Gaussians imply that all the described distributions are Gaussian, and then to use simple formulas to determine the parameters of those distributions.

(a) We can construct Y via $Y = X + Z$ where $Z \sim \mathcal{N}(0, \Sigma')$ is independent of X. Since sums of Gaussians are Gaussian, this shows that Y is normally distributed, which means that we only need to specify the mean and covariance matrix for the distribution. To find its mean, we use the law of total expectation to write

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(X) = \mu.$$

Or, alternatively, linearity of expectation:

$$\mathbb{E}(Y) = \mathbb{E}(X + Z) = \mu + 0 = \mu.$$

To find the variance of Y, we can use the law of total variance (a.k.a. Eve's law), which gives

$$\text{var}(Y) = \text{var}(\mathbb{E}(Y|X)) + \mathbb{E}(\text{var}(Y|X)) = \text{var}(X) + \mathbb{E}(\Sigma') = \Sigma + \Sigma'.$$

Or, alternatively:

$$\text{var}(Y) = \text{var}(X + Z) = \text{var}(X) + \text{var}(Z) + \text{cov}(X, Z) = \Sigma + \Sigma' + 0 = \Sigma + \Sigma'.$$

Check 3.1: You showed the the unconditional distribution of Y is normal (either by the method above, where we use $Y - X$, or by multiplying the density of $Y|X$ and X, and observing that the joint distribution is MVN, so the marginal distribution of Y must also be).

Check 3.2: You found that the mean was μ, the covariance matrix was $\Sigma + \Sigma'$, and explained that that suffices to specify the distribution.

(b) Since X, Y, and $Y|X$ are all Gaussian, then (X, Y) is Gaussian as well (see note below), and so we need only to compute its mean and variance. We already know its mean: it is (μ, μ). What about the covariance? It will be a $2m \times 2m$ matrix, and its top-left $m \times m$ block will just be the variance of X, i.e., Σ. Similarly its bottom-right $m \times m$ block will be the variance of Y, i.e., $\Sigma + \Sigma'$. It therefore remains

only to compute its off-diagonal block, which is the covariance of X and Y. This can be computed using the law of total covariance, which gives that:

$$\begin{aligned}\text{cov}(X, Y) &= \text{cov}(\mathbb{E}(X|X), \mathbb{E}(Y|X)) + \mathbb{E}(\text{cov}(X, Y|X)) \\ &= \text{cov}(X, X) + 0 \\ &= \Sigma.\end{aligned}$$

Or, alternatively, you can write $Y = X + Z$ where $Z \sim \mathcal{N}(0, \Sigma')$ is independent of X, and then use bilinearity of covariance to obtain $\text{cov}(X, X + Z) = \text{cov}(X, X) + \text{cov}(X, Z) = \Sigma$.

These computations give the following solution:

$$(X, Y) \sim \mathcal{N}\left((\mu, \mu), \begin{bmatrix}\Sigma & \Sigma \\ \Sigma & \Sigma + \Sigma'\end{bmatrix}\right)$$

Note that ordinarily we would have to take care to make one of the off-diagonal blocks of the above covariance matrix the transpose of the other, but here $\Sigma^T = \Sigma$.

Note: it may not be clear that given $X, Y, Z = Y - X$ are all Gaussian, then (X, Y) also is. Note that we do need $Y|X$ to be Gaussian (see the example on Wikipedia). Remember that a vector X is a multivariate normal iff it can be written as $X = \mu + A\mathbf{m}$, where $\mathbf{m}$ is a vector of i.i.d. 1-dimensional normals. Then, letting $Z = A'\mathbf{n}$ in the same way, we have:

$$\begin{bmatrix}X \\ Y\end{bmatrix} = \begin{bmatrix}X \\ X + Z\end{bmatrix} = \begin{bmatrix}\mu + A\mathbf{m} \\ \mu + A\mathbf{m} + A'\mathbf{n}\end{bmatrix} = \begin{bmatrix}\mu \\ \mu\end{bmatrix} + \begin{bmatrix}A & 0 \\ A & A'\end{bmatrix}\begin{bmatrix}\mathbf{m} \\ \mathbf{n}\end{bmatrix}$$

In fact, we can actually finish the problem from here, since the matrix A is the Cholesky decomposition of the covariance matrix of the multivariate normal, so the covariance is:

$$\begin{bmatrix}A & 0 \\ A & A'\end{bmatrix}\begin{bmatrix}A & 0 \\ A & A'\end{bmatrix}^\intercal = \begin{bmatrix}AA^\intercal & AA^\intercal \\ AA^\intercal & A'A'^\intercal + AA^\intercal\end{bmatrix} = \begin{bmatrix}\Sigma & \Sigma \\ \Sigma & \Sigma + \Sigma'\end{bmatrix}$$

Check 3.3: You justified that the joint distribution was multivariate normal, either by taking the product of the densities as above, or by finding the matrix that takes i.i.d. Gaussians to the vector (X, Y). There are other solutions possible.

Check 3.4: You used either the law of total expectation and law of total covariance or found the transformation of i.i.d. Gaussians and used the Cholesky decomposition to find the mean and covariance matrices.

I can Ei-gen

Problem 4

Let $\mathbf{X} \in \mathbb{R}^{n \times m}$.

(a) What is the relationship between the n eigenvalues of $\mathbf{X}\mathbf{X}^T$ and the m eigenvalues of $\mathbf{X}^T\mathbf{X}$?

(b) Suppose $\mathbf{X}$ is square (i.e., $n = m$) and symmetric. What does this tell you about the eigenvalues of $\mathbf{X}$? What are the eigenvalues of $\mathbf{X} + \mathbf{I}$, where $\mathbf{I}$ is the identity matrix?

(c) Suppose $\mathbf{X}$ is square, symmetric, and invertible. What are the eigenvalues of $\mathbf{X}^{-1}$?

Hints:

- Make use of singular value decomposition and the properties of orthogonal matrices. Show your work.
- Review and make use of (but do not derive) the spectral theorem.

Solution

(a) The non-zero eigenvalues are all the same. This can be seen by using the singular value decomposition to write $X = U\Sigma V^T$ where Σ is an $n \times m$ diagonal matrix and U and V are unitary matrices. Then

$$XX^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T$$

and

$$X^T X = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T.$$

Suppose without loss of generality that $n \leq m$, so that $d_1, \ldots, d_n$ are the diagonal entries of Σ. Then since Σ is diagonal it is easy to verify that $\Sigma\Sigma^T$ is an $n \times n$ matrix with diagonal entries $d_1^2, \ldots, d_n^2$, and that $\Sigma^T\Sigma$ is a diagonal matrix whose first n diagonal elements are $d_1^2, \ldots, d_n^2$ while the rest are 0.

Check 4.1: You noticed that the non-zero eigenvalues are the same, and showed it briefly.

(b) If X is square and symmetric then the spectral theorem tells us that the eigenvalues of X are all real. To analyze $X = I$, we can write X as $U\Sigma U^T$ where U is a unitary matrix whose columns are the eigenvectors of X and Σ is a diagonal matrix whose diagonal entries are the eigenvalues $\lambda_1, \ldots, \lambda_n$ of X. Since $I = UIU^T$, this means that $X + I = U(\Sigma + I)U^T$. Therefore, the i-th eigenvalue of $X + I$ is $\lambda_i + 1$.

Another way is to use the alternative definition of eigenvalues as the scaling that happens to eigenvectors, and note that X and $X + I$ share the same eigenvectors.

Check 4.2: You noted that the spectral theorem gives us that the eigenvalues are real.

Check 4.3: Either you noticed that X being symmetric and square gives us that the two orthonormal transformations in the SVD are the same, and used that fact to show that the eigenvalues are 1 more, or you used the alternative definition of eigenvalues to get the same fact.

(c) Using the same representation as above, i.e., $X = U\Sigma U^T$, the condition that X be invertible is equivalent to all of the eigenvalues λ_i being non-zero. Therefore, Σ is invertible, and its inverse is a diagonal matrix whose i-th diagonal entry is $1/\lambda_i$. It is then easy to see that $X^{-1} = U\Sigma^{-1}U^T$, for

$$(U\Sigma^{-1}U^T)X = U\Sigma^{-1}U^T U\Sigma U^T = U\Sigma\Sigma^{-1}U^T = UU^T = I.$$

This means that the eigenvalues of X^{-1} are the diagonal entries of Σ^{-1}, namely $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$.

Check 4.4: Using some method (probably similar to the one above) you showed that the eigenvalues are the inverse of the eigenvalues of X.

Vector Calculus

Problem 5

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ and $\mathbf{A} \in \mathbb{R}^{m \times m}$. Please derive from elementary scalar calculus the following useful properties. Write your final answers in vector notation.

(a) What is the gradient with respect to $\mathbf{x}$ of $\mathbf{x}^T\mathbf{y}$?

(b) What is the gradient with respect to $\mathbf{x}$ of $\mathbf{x}^T\mathbf{x}$?

(c) What is the gradient with respect to $\mathbf{x}$ of $\mathbf{x}^T\mathbf{A}\mathbf{x}$?

Solution

(a) $\partial(x^T y)/\partial x_i = \partial(\sum_i x_i y_i)/\partial x_i = y_i$, so $\nabla(x^T y) = y$.

Check 5.1: You correctly derived that the solution was y. A solution of $y^\top$ is fine for now.

(b) $\partial(x^T x)/\partial x_i = \partial(\sum_i x_i^2)/\partial x_i = 2x_i$, so $\nabla(x^T x) = 2x$.

Check 5.2: You correctly derived that the solution was $2x$. A solution of $2x^T$ is fine for now, but $x + x^T$ is not.

(c) This one is a bit more involved. We write

$$\begin{aligned}
\frac{\partial}{\partial x_i}(x^T A x) &= \frac{\partial}{\partial x_i} \sum_{j,k} A_{jk} x_j x_k \\
&= \frac{\partial}{\partial x_i}\Big(\sum_{j,k\neq i} A_{jk} x_j x_k + \sum_{k\neq i} A_{ik} x_i x_k + \sum_{j\neq i} A_{ji} x_j x_i + A_{ii} x_i^2\Big) \\
&= \sum_{k\neq i} A_{ik} x_k + \sum_{j\neq i} A_{ji} x_j + 2A_{ii} x_i \\
&= \sum_{k} A_{ik} x_k + \sum_{j} A_{ji} x_j \\
&= (Ax)_i + (x^T A)_i
\end{aligned}$$

which means that $\nabla(x^T A x) = (A + A^T)x$.

Check 5.3: You correctly derived that the solution was $(A + A^T)x$. A solution of $2Ax$ is not correct.

Gradient Check

Problem 6

Often after finishing an analytic derivation of a gradient, you will need to implement it in code. However, there may be mistakes - either in the derivation or in the implementation. This is particularly the case for gradients of multivariate functions.

One way to check your work is to numerically estimate the gradient and check it on a variety of inputs. For this problem we consider the simplest case of a univariate function and its derivative. For example, consider a function $f(x) : \mathbb{R} \to \mathbb{R}$:

$$\frac{df}{dx} = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x-\epsilon)}{2\epsilon}$$

A common check is to evaluate the right-hand side for a small value of ϵ, and check that the result is similar to your analytic result.

In this problem, you will implement the analytic and numerical derivatives of the function

$$f(x) = \cos(x) + x^2 + e^x.$$

1. Implement `f` in Python (feel free to use whatever `numpy` or `scipy` functions you need):

   ```
   def f(x):
   ```

2. Analytically derive the derivative of that function, and implement it in Python:

   ```
   def grad_f(x):
   ```

3. Now, implement a gradient check (the numerical approximation to the derivative), and by plotting, show that the numerical approximation approaches the analytic as `epsilon` $\to 0$ for a few values of x:

   ```
   def grad_check(x, epsilon):
   ```

Solution

1. Let $f(x) = \cos(x) + x^2 + e^x$.

   ```
   import numpy as np
   def f(x):
       return np.cos(x) + x**2 + np.exp(x)
   ```

   Check 6.1: You correctly expressed the function in code in Python or in another language.

2. The derivative $f'(x)$ is given by

   $$f'(x) = -\sin(x) + 2x + e^x$$

   ```
   import numpy as np
   def grad_f(x):
       return np.sin(x) + 2*x + np.exp(x)
   ```

   Check 6.2 Your derivative is correct.

3. The answer is therefore

```
import numpy as np
import matplotlib.pyplot as plt
def grad_check(x, epsilon):
    return (f(x + epsilon) - f(x - epsilon))/(2*epsilon)
e = np.linspace(0.0001, 1, 100)
plt.plot(e, grad_check(0, e))  # check the gradient at x=0
plt.show()
```

Check 6.3 Your approximation and analaytic solution are close for small values of epsilon and further for larger epsilon. You plotted your results to show the residuals between the approximation and analytic solution.