Aditya Prasad
adityaprasad@college.harvard.edu
CS281-F17

Assignment #0, v 1.0
Due: 5:00pm Sept. 8th

Collaborators: None

Homework 0: Preliminary

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, and have studied the mathematical review notes provided in section. 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:

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

(a) $X \perp Y$. Therefore, we have:

$$Cov(X, Y) = E((X - E(X))(Y - E(Y)) \tag{1}$$
$$= E(X - E(X))E(Y - E(Y)) \tag{2}$$
$$= 0 \tag{3}$$

Alternatively, we have:

$$Cov(X, Y) = E(XY) - E(X)E(Y) - E(XE(Y)) + E(YE(X)) \tag{4}$$
$$= 2E(X)E(Y) - 2E(X)E(Y) = 0 \tag{5}$$

Here, we have assumed the general result that if two random variables are independent, the expectation of the product is then the product of the expectation.

(b) Covariance is a measure of linear association; therefore, consider some solely non-linear dependence: $Y = X^2, X \sim N(0, 1)$. To heuristically see why there isn't any covariance, consider moving across the possible values of X from $-\infty$ to ∞. First an increase in X is associated with a decrease in Y. But after crossing $X = 0$, an increase in X is associated with an increase in Y. Overall, however, these two effects essentially cancel out.

$$Cov(X, Y) = E(XY) - E(X)E(Y) \tag{6}$$
$$= E(X^3) - E(X)E(X^2) \tag{7}$$
$$= 0 - 0 = 0 \tag{8}$$

Y and X are obviously not independent, but they are uncorrelated.

(c) Consider two random variable X and Y. Here, we assume that they are continuous. The pdf of X is f_X, and that of Y is f_Y. Therefore, $X + aY$ is distributed as $f_{X,Y}$ (the joint pdf of X and Y). Therefore, we have:

$$E(X + aY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x + ay)f_{X,Y}(x, y)dxdy \tag{9}$$
$$= \int_{-\infty}^{\infty} xdx \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy + a\int_{-\infty}^{\infty} ydy \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx \tag{10}$$

We have put this in a form where we can simply integrate out the unnecessary terms and move from the joint pdfs to marginal pdfs.

$$E(X + aY) = \int_{-\infty}^{\infty} x f_X(x) dx + a \int_{-\infty}^{\infty} y f_Y(y) dy \tag{11}$$

$$= E(X) + aE(Y) \tag{12}$$

Note, this theorem is true irrespective of whether X and Y are independent. Now, we prove the corresponding theorem for variance, where this is not true.

$$Var(X + aY) = E((X + aY)^2) - E^2(X + aY) \tag{13}$$

$$= E(X^2 + a^2Y^2 + 2aXY) - (E(X) + aE(Y))^2 \tag{14}$$

$$= E(X^2) - E^2(X) + a^2(E(Y^2) - E^2(Y)) + 2aE(XY) - 2aE(X)E(Y) \tag{15}$$

$$= Var(X) + a^2 Var(Y) + 2aCov(X, Y) \tag{16}$$

If $X \perp Y$, then we have:

$$Var(X + aY) = Var(X) + a^2 Var(Y) \tag{17}$$

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.

(a) Yes. The only constraint on a probability density function is that it must integrate to 1. An obvious display of this is the following uniform distribution, $f_X(x) = 2, x \in [0, 1/2]$. A neater description of this is the Dirac-delta distribution, which essentially represents certainty about an event.

(b) The pdf is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{x^2}{\sigma^2}\right] \tag{18}$$

where $\sigma^2 = 1/100$.

(c) $f_X(0) = \frac{1}{\sqrt{2\pi}\sigma} = \frac{10}{\sqrt{2\pi}}$

(d) $P(X = 0) = \int_0^0 f_X(x)dx = 0$

(e) The set $X = 0$ is of volume 0; therefore, the probability that it happens is zero (given that we're integrating over a set of volume zero). More intuitively, given that there an un-countably infinite number of reals between $X = -\infty \to \infty$, the probability that X equals exactly 0 is 0.

A better framed-question would to be ask what the probability of $X = 0$ is around a small ball of radius $\epsilon > 0$ centered at $X = 0$. We then have:

$$P(X = 0) = -\int_{0-\epsilon}^{0+\epsilon} f_X(x)dx \tag{19}$$

$$= \frac{10}{\sqrt{2\pi}} erf(\epsilon) \tag{20}$$

As can be seen, as $\epsilon \to 0, P(X = 0) \to 0$.

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.

(a) Given that $Y \mid X$ and X are distributed as Gaussians, from the closure properties of normals, we know that the marginal distribution of Y should also be Gaussian. Therefore, we only need to calculate the mean and variance to fully specify this distribution.

From Adam's law, we have:

$$E(Y) = E(E(Y|X)) = E(X) = \mu \in \mathbb{R}^m \tag{21}$$

From Eve's law, we have:

$$\begin{aligned} Var(Y) &= Var(E(Y|X)) + E(Var(Y|X)) && (22)\\ &= Var(X) + \mathbf{\Sigma}^* && (23)\\ &= \mathbf{\Sigma} + \mathbf{\Sigma}^* && (24) \end{aligned}$$

(b) Now we calculate the joint distribution of the vector: $(X, Y)^T$. Given that the marginals are normal, from the closure properties of normals, we know that joint distribution is also distributed as a normal. Therefore, we need to calculate only the mean vector and covariance matrix to specify this distribution. The mean is simply going to be: $(\mu, \mu) \in \mathbb{R}^{2m}$. The quadratic form looks like:

$$(\mathbf{x}, \mathbf{y})\Sigma_{XY}^{-1}(\mathbf{x}, \mathbf{y})^T \tag{25}$$

We know that the covariance matrix for such a system looks like:

$$\begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(Y,X) & Var(Y) \end{pmatrix}$$

We know $var(X) = \Sigma$ and $var(Y) = \Sigma + \Sigma^*$. We also know that $Cov(X, Y) = Cov(Y, X)$. We now calculate $Cov(X, Y)$. From the law of total covariance, we have:

$$Cov(X, Y) = E(Cov(X, Y|Z)) + Cov(E(X|Z), E(Y|Z)) \tag{26}$$

Let $Z = X$. Then, we have:

$$\begin{aligned} Cov(X, Y) &= E(Cov(X, Y|X)) + Cov(X, X) && (27)\\ &= 0 + Var(X) && (28) \end{aligned}$$

Another way to do calculate $Cov(X, Y)$ is through the definition:

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) && (29)\\ &= E_X\left[XE_{Y|X}[Y]\right] - E(X)E(Y) && (30)\\ &= E_X[X^2] - E(X)E(X) && (31)\\ &= Var(X) && (32) \end{aligned}$$

Also, because of symmetry: $Cov(Y, X) = Cov(X, Y)$. Therefore, we have fully specified the matrix:

$$\begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} \\ \mathbf{\Sigma} & \mathbf{\Sigma} + \mathbf{\Sigma}' \end{pmatrix}$$

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.

(a) Any $X \in \mathbb{R}^{m \times n}$ admits a singular value decomposition. Thus, we can write X as follows:

$$X = \Sigma U V^T \tag{33}$$

Here, Σ and V are orthogonal matrices, while U is a diagonal matrix. Now, consider XX^T. We write this out using the singular value decomposition:

$$XX^T = (\Sigma U V^T)(V U^T \Sigma^T) \tag{34}$$
$$= \Sigma U (V^T V) U^T \Sigma^T \tag{35}$$
$$= \Sigma U U^T \Sigma^T \tag{36}$$

There are a couple of important things to note: firstly, XX^T is obviously a square symmetric matrix. ($(XX^T)^T = XX^T$). Since (by the spectral theorem for symmetric matrices) every symmetric matrix admit admits a unique eigenvalue decomposition (up to a constant) of the form:

$$XX^T = P^{-1} D P \tag{37}$$

Here, P is an orthogonal eigenbasis matrix, while D is the diagonal eigenvalue matrix (essentially principal axis theorem). Therefore on comparing the two forms above, we can conclude that:

$$D = UU^T \tag{38}$$
$$P^{-1} = \Sigma \tag{39}$$

Therefore, the eigenvalues of XX^T are stored in the diagonal matrix UU^T (the product of two diagonal matrices is another diagonal matrix). The second thing to note here is that because U is diagonal, it is symmetric; that is $U = U^T$. Therefore,

$$UU^T = U^2 \tag{40}$$
$$\implies D = U^2 \tag{41}$$

We can do a similar spectral decomposition for $X^T X$, and use eigen-decomposition to see its matrix of eigenvalues.

$$X^T X = V U^T \Sigma^T \Sigma U V^T \tag{42}$$
$$= V U^T U V^T \tag{43}$$
$$= V U^2 \Sigma \tag{44}$$

By the same eigenvalue-decomposition argument as above, we note that the diagonal eigenvalue matrix, $D^{'}$ of $X^T X$ is:

$$D^{'} = U^2 \tag{45}$$

Therefore, we note that the eigenvalues of the two matrices are the same. The eigenvectors, however, are generally different. The fact that the eigenvalues are the same seems to relate to the fact that the row rank is the same as the column rank. In other words, the dimension of both spaces is the same. We have essentially constructed an orthonormal basis in either space, and these bases seem to have a similar "geometric shape" (evinced by their eigenvalues).

(b) If $X \in GL_n(\mathbb{R})$ is symmetric, we know (by the Spectral Theorem) that the eigenvalues of X are real. The eigenvalues of $(X + I)$ are essentially each of the eigenvalues plus one. Let v be an eigenvector if X. Then:

$$(X + I)v = (\lambda + 1)v \tag{46}$$

(c) $X = X^T$. Let $\mathbf{v}$ be an eigenvector of X with eigenvalue λ. Then:

$$X\mathbf{v} = \lambda\mathbf{v} \tag{47}$$

$$\implies X^{-1}X\mathbf{v} = \lambda X^{-1}\mathbf{v} \tag{48}$$

$$\implies X^{-1}\mathbf{v} = \frac{1}{\lambda}\mathbf{v} \tag{49}$$

Therefore, the eigenvalues are each the multiplicate inverse of those of the forward function. This is independent of whether X is symmetric. In essence, the forward function scales each eigenvector, and so the inverse function just scales them back.

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 $\mathbf{x}^T\mathbf{A}\mathbf{x}$?

(a) Consider the inner product: $\langle \mathbf{x}, \mathbf{y} \rangle$. In Einstein notation (two same indices indicate a summation), we have:

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_i y_i \tag{50}$$

The derivate with respect to x_j is simply:

$$\partial_{x_j} \langle \mathbf{x}, \mathbf{y} \rangle = \partial_{x_j} x_i y_i \tag{51}$$
$$= y_i \delta_{ij} = y_j \tag{52}$$

Therefore, the gradient with respect to $\mathbf{x}$ of $\mathbf{x}^T\mathbf{y}$ is simply $\mathbf{y}$.

(b) Similarly, we have:

$$\langle \mathbf{x}, \mathbf{x} \rangle = x_i x_i \tag{53}$$

Therefore,

$$\partial_{x_j} \langle \mathbf{x}, \mathbf{x} \rangle = \partial_{x_j} x_i x_i \tag{54}$$
$$= 2x_i \delta_{ij} = 2x_j \tag{55}$$

Therefore, the gradient with respect to $\mathbf{x}$ of $\mathbf{x}^T\mathbf{y}$ is simply $2\mathbf{x}$.

(c) Assume now that the inner product space is given by the matrix A, we have:

$$\langle \mathbf{x}, \mathbf{x} \rangle = x_i A_{ij} x_j \tag{56}$$

Therefore,

$$\partial_{x_k} \langle \mathbf{x}, \mathbf{x} \rangle = \partial_{x_k} x_i A_{ij} x_j \tag{57}$$
$$= [A_{ij} x_j] \, \delta_{ik} + [A_{ij} x_i] \, \delta_{jk} \tag{58}$$
$$= A_{kj} x_j + A_{ik} x_i \tag{59}$$
$$= A_{kj} x_j + A^T_{ki} x_i \tag{60}$$
$$\tag{61}$$

Therefore, in vector notation, the gradient is:

$$(\mathbf{A} + \mathbf{A^T})\mathbf{x} \tag{62}$$

It makes sense that the matrix representing the gradient operator in this space is symmetric; the angle of steepest ascent for $\mathbf{e}_i$, $\mathbf{e}_j$ should be the same for $\mathbf{e}_j$, $\mathbf{e}_i$.

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):
   ```

```
import numpy as np
import matplotlib.pyplot as plt
```

(a)
```
def f(x):
    return np.cos(x) + x**2 + np.exp(x)
```

(b)
```
def grad_f(x):
    return -np.sin(x) + 2.*x + np.exp(x)
```

(c)
```
def grad_check(x, eps):
    return (f(x + eps) - f(x-eps)) / float((2*eps))

def plot():
    xs = [1., 5., 6., 8.,10.]
    eps = np.linspace(0.001,1, 100)
    vect = np.vectorize(grad_check)
```

```
cnt = 0
plt.figure(figsize=(8,8))
for x in xs:
    cnt = cnt + 1
    print(cnt)
    plt.xlabel("Epsilon")
    plt.ylabel("Error")
    y = vect(x,eps) - grad_f(x)
    plt.plot(eps, y, label="x = %2.2f"%x)
    plt.legend()
plt.savefig('plot.pdf', dpi=600, bbox_inches='tight', transparent=False)
```

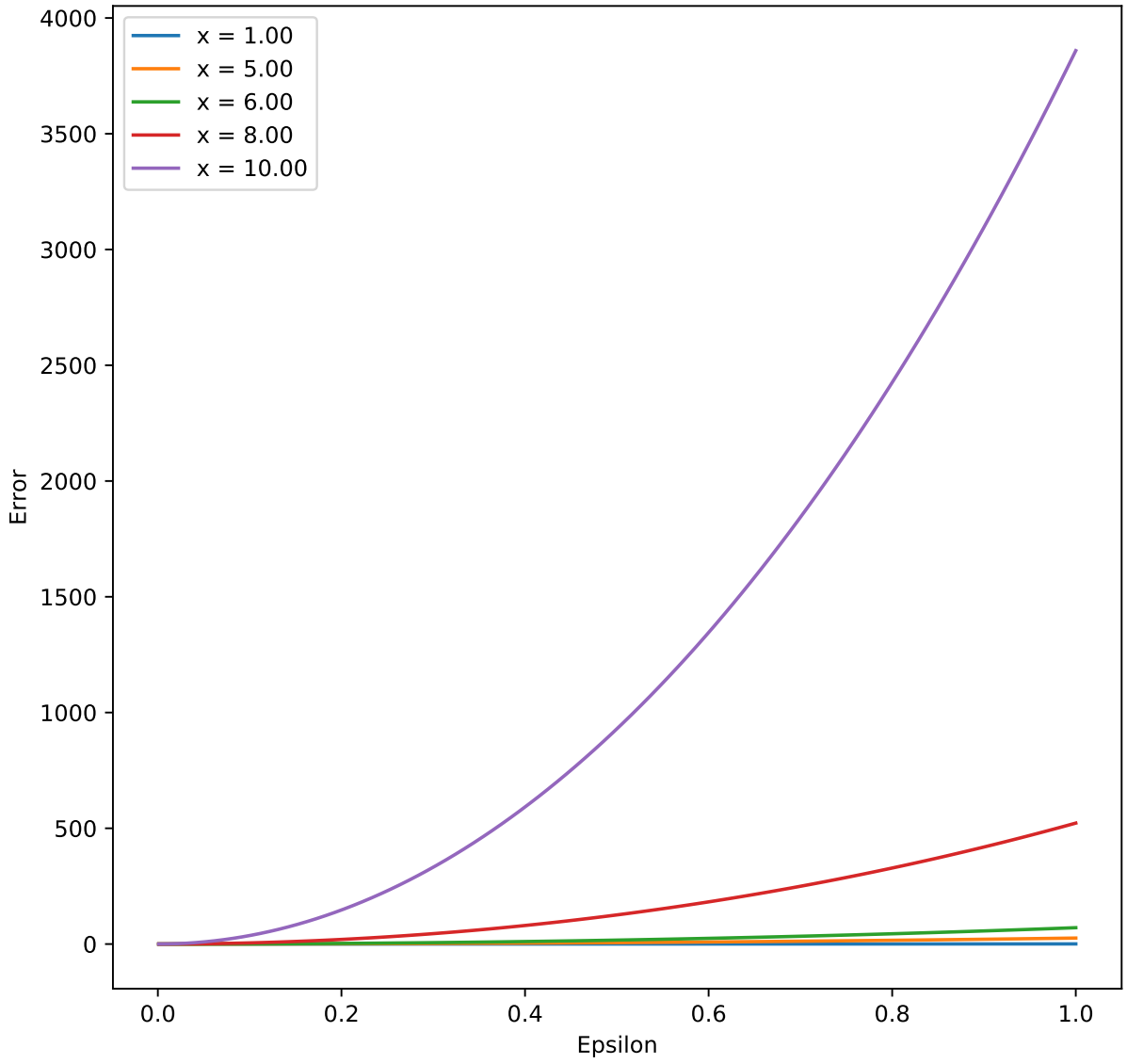

Fig 1. Error values as a function of epsilon for different values of x