Melissa Yu
melissayu@college.harvard.edu
CS281-F17

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

Collaborators: Alex Lin

Homework 0: Preliminary

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, 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:

$$\begin{aligned}\mathbb{E}(X + aY) &= \mathbb{E}(X) + a\mathbb{E}(Y) \\ \mathrm{var}(X + aY) &= \mathrm{var}(X) + a^2\mathrm{var}(Y)\end{aligned}$$

(a) As shown in math review, we have $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ for independent variables X and Y. Expanding the expression for covariance, we obtain the desired result:

$$\begin{aligned}\mathrm{cov}(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] \\ &= \mathbb{E}\left[(XY - \mathbb{E}(X)Y - \mathbb{E}(Y)X + \mathbb{E}(X)\mathbb{E}(Y)\right] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) && \text{by linearity of expectation} \\ &= 0\end{aligned}$$

(b) Take $X \sim \mathcal{N}(0, 1)$ and $Y \sim X^2$. Then, we have the covariance

$$\begin{aligned}\mathrm{cov}(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}(X))(X^2 - \mathbb{E}(X^2))\right] \\ &= \mathbb{E}(X^3) \\ &= 0,\end{aligned}$$

where $\mathbb{E}(X^3)$ can be shown to be 0 by evaluating the third derivative of the moment generating function for the standard normal at 0.

(c) For the case where X and Y are discrete R.V.s, we have:

$$\begin{aligned}\mathbb{E}(X + aY) &= \sum_{x,y}(x + ay)P(X = x, Y = y) \\ &= \sum_{x,y} xP(X = x, Y = y) + \sum_{x,y} ayP(X = x, Y = y) \\ &= \sum_{x} xP(X = x) + \sum_{y} ayP(Y = y) \\ &= \mathbb{E}(X) + a\mathbb{E}(Y)\end{aligned}$$

The proof for continuous R.V.s uses integrals instead of summations, and is analogous.

$$\begin{aligned}\mathrm{var}(X + aY) &= \mathbb{E}\left[(X + aY)^2\right] - \mathbb{E}\left[X + aY\right]^2 \\ &= \mathbb{E}\left[X^2 + 2aXY + a^2Y^2\right] - \left(\mathbb{E}(X)^2 + 2a\mathbb{E}(X)\mathbb{E}(Y) + a^2\mathbb{E}(Y)^2\right) \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 + a^2(\mathbb{E}(Y^2) - \mathbb{E}(Y)^2) \\ &= \mathrm{var}(X) + a^2\mathrm{var}(Y)\end{aligned}$$

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 pdf is just a measure of the *relative* probability of a value.

(c) Substituting into the pdf for the normal, we have:

$$X \sim \mathcal{N}(0, 1/100) \implies f_X(x) = \frac{10}{\sqrt{2\pi}} \exp\left(-50x^2\right)$$

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

(d) $P(X = 0) = 0$

(e) The real probability of a continuous R.V. equaling an exact value is 0 (there are infinitely many values on the support), but the relative probability of this event is not 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) From the closure properties for a multinomial distribution, we have that Y is distributed MVN with some mean and covariance matrix. To completely specify the distribution, we need only compute these two parameters:

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}(X) = \mu \\ \operatorname{cov}(Y, Y) &= \mathbb{E}(\operatorname{var}(Y \mid X)) + \operatorname{var}(\mathbb{E}(Y \mid X)) \\ &= \mathbb{E}(\mathbf{\Sigma}') + \operatorname{var}(X) \\ &= \mathbf{\Sigma}' + \mathbf{\Sigma}\end{aligned}$$

Thus,

$$Y \sim \mathcal{N}(\mu, \mathbf{\Sigma} + \mathbf{\Sigma}')$$

(b) We can understand this joint distribution as the vector $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$. As before, we know this is distributed MVN by the closure properties. Then, the expected value is simply

$$\mathbb{E}(Z) = \begin{bmatrix} \mathbb{E}(X) \\ \mathbb{E}(Y) \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix}$$

Similarly, the covariance matrix of Z can be expressed as:

$$\operatorname{cov}(Z, Z) = \begin{bmatrix} \operatorname{cov}(X, X) & \operatorname{cov}(X, Y) \\ \operatorname{cov}(Y, X) & \operatorname{cov}(Y, Y) \end{bmatrix}$$

We can use the law of total covariance to compute the off-diagonal covariances:

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathbb{E}(\operatorname{cov}(X, Y \mid X)) + \operatorname{cov}(\mathbb{E}(X \mid X), \mathbb{E}(Y \mid X)) \\ &= 0 + \operatorname{cov}(X, X),\end{aligned}$$

where the covariance of X and Y given X is 0, since the two are independent given X. Thus, we have the joint distribution

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

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) Let the SVD of $\mathbf{X}$ be $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{\Sigma}$ is diagonal. Making use of the fact that $\mathbf{V}\mathbf{V}^T = \mathbf{I}$ for any orthogonal matrix $\mathbf{V}$, we have the following diagonalizations:

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T$$
$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T$$

Since $\mathbf{\Sigma}$ is a diagonal matrix, $\mathbf{\Sigma}^T\mathbf{\Sigma}$ and $\mathbf{\Sigma}\mathbf{\Sigma}^T$ have the same values on their diagonals, except for zero padding. Thus, we conclude that the two matrices have the same non-zero eigenvalues, with the larger matrix having additional zero-eigenvalues.

(b) By the spectral theorem, the eigenvalues of $\mathbf{X}$ are all real and correspond to n orthogonal eigenvectors. For any eigenvalue λ with eigenvector v of $\mathbf{X}$, we have $\mathbf{X}v = \lambda v \implies (\mathbf{X} + \mathbf{I})v = (\lambda + 1)v$. Thus, the eigenvalues of the new matrix are found by adding one to the eigenvalues of the original matrix.

(c) For any eigenvalue λ with eigenvector v of $\mathbf{X}$, we have

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

Thus, the eigenvalues of $\mathbf{X}^{-1}$ are simply the reciprocals of the original eigenvalues.

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) First, we observe that

$$\frac{\partial \mathbf{x}^T\mathbf{y}}{\partial x_k} = \frac{\partial \sum_{i=1}^m x_i y_i}{\partial x_k} = y_k$$

Combining, we obtain

$$\frac{\partial \mathbf{x}^T\mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{x}^T\mathbf{y}}{\partial x_1} \\ \vdots \\ \frac{\partial \mathbf{x}^T\mathbf{y}}{\partial x_m} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \mathbf{y}$$

(b) Again, we expand the dot product as

$$\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial x_k} = \frac{\partial \sum_{i=1}^m x_i^2}{\partial x_k} = 2x_k$$

Thus,

$$\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{x}^T\mathbf{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \mathbf{x}^T\mathbf{x}}{\partial x_m} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_m \end{bmatrix} = 2\mathbf{x}$$

(c) Noting that $\mathbf{x}^T\mathbf{A}\mathbf{x} = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^m \sum_{j=1}^m x_i A_{ij} x_j$, we have:

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial x_k} = \sum_{j=1}^m (A_{jk} + A_{kj}) x_j$$

As before, we obtain:

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial x_1} \\ \vdots \\ \frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial x_m} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^m (A_{j1} + A_{1j}) x_j \\ \vdots \\ \sum_{j=1}^m (A_{jm} + A_{mj}) x_j \end{bmatrix} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$$

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 $f(x) : \mathbb{R} \rightarrow \mathbb{R}$:

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

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

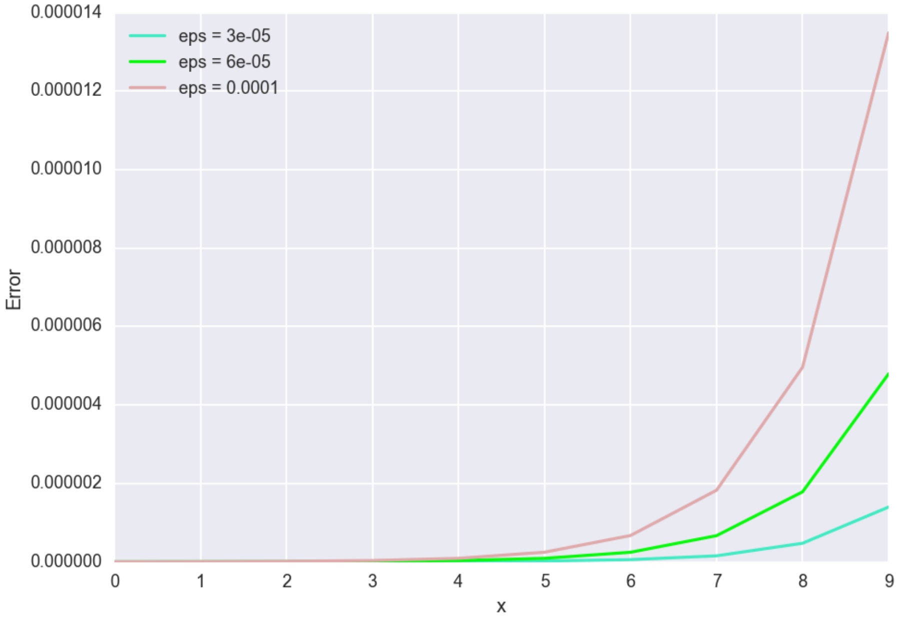
2. Analytically derive the derivative of that function, and implement it in Python:

```
def grad_f(x):
        return - np.sin(x) + 2 * x + np.exp(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 $\rightarrow 0$ for a few values of x:

```
def grad_check(x, epsilon):
        return (f(x + epsilon) - f(x - epsilon)) / (2 * epsilon)
```

The function implementations are completed above, and can also be found in the python files submitted to Canvas. The graph below shows the difference between the analytic and numerical solutions for various epsilon. As expected, the difference becomes smaller as epsilon goes to 0.


eps = 3e-05
eps = 6e-05
eps = 0.0001
Error
x