Ben Cook
bcook@cfa.harvard.edu
CS281-F17

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

Collaborators: Roxana Pop

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:

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

Variance and Covariance – Solution

(a) The covariance is defined to be:

$$\text{cov}(X, Y) \equiv \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

If X and Y are independent, then $p(x, y) = p(x)p(y)$, and the expectation of them can be separated:

$$\begin{aligned}\mathbb{E}(XY) &\equiv \int\int xyp(x, y)dxdy = \int\int xyp(x)p(y)dxdy \\ &= \int xp(x)dx \int yp(y)dx \\ &= \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

Therefore, the covariance is necessarily zero.

(b) As one example, let us take $X \sim \mathcal{N}(0, 1)$, and $Y = X^2$. X and Y are clearly highly dependent, but we can calculate their covariance:

$$\begin{aligned}\text{cov}(X, Y) &\equiv \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2)\end{aligned}$$

From the fact that $p(x)$ is symmetric about zero, we can see that $\mathbb{E}(X)$ and $\mathbb{E}(X^3)$ are both zero. Therefore, without even needing to calculate $\mathbb{E}(X^2)$ we can find:

$$\boxed{\text{cov}(X, Y) = 0 - 0\mathbb{E}(X^2) = 0}$$

Therefore, we have found two random variables which are not independent but do have zero covariance.

(c)

$$\begin{aligned}
\mathbb{E}(X+aY) &\equiv \int\int (x+ay)p(x)p(y)dxdy \\
&= \left(\int xp(x)dx\right) + a\left(\int yp(y)dy\right) \\
\therefore \boxed{\mathbb{E}(X+aY) = \mathbb{E}(X) + a\mathbb{E}(Y)} & \\
\mathrm{var}(X+aY) &\equiv \mathbb{E}\left[(X+aY)^2\right] - \mathbb{E}\left[X+aY\right]^2 \\
&= \mathbb{E}\left[X^2 + a^2Y^2 + 2aXY\right] - \mathbb{E}\left[X+aY\right]^2
\end{aligned}$$

From the independence of X and Y and linearity of expectation:

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

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.

Densities – Solutions

(a) Yes, the only relevant restriction on a pdf is that it *integrates to one* and is never negative. Therefore, the pdf can take on values greater than 1, but just not over very large portions of its domain. For example:

$$p(x) = \begin{cases} 10\,, & 0 \leq x < 0.1 \\ 0\,, & \text{else} \end{cases}$$

is a perfectly valid, normalized pdf which takes on values greater than 1.

(b) The pdf of a univariate normal distribution is:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Plugging in the values provided:

$$p\left(x|\mu = 0, \sigma^2 = 0.01\right) = \frac{10}{\sqrt{2\pi}} \exp\left[-50x^2\right],$$

defined for all real values of x.

(c)

$$\begin{aligned} p\left(x = 0|\mu = 0, \sigma^2 = 0.01\right) &= \frac{10}{\sqrt{2\pi}} \\ &\approx 3.99 \end{aligned}$$

(d) The probability that X is exactly equal to 0 is zero:

$$P(X = 0) = \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} p(x')dx' = 0$$

(e) While the pdf represents the probability *mass* ($p(x)$) at a particular point of its domain, this mass must be integrated over a particular segment of the domain (dx) to compute the probability of the random variable X lying in that segment of the domain. There is zero probability that a *continuous* variable will ever be *exactly* equal to zero, or any specific number for that matter.

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.

Conditioning and Bayes' rule – Solutions

1. We can compute the unconditional (a.k.a. marginal) probability by marginalizing over X:

$$\begin{aligned} p(y) &= \int p(y|x)p(x)dx = \int \mathcal{N}(y|x, \mathbf{\Sigma}')\mathcal{N}(x|\mu, \mathbf{\Sigma})dx \\ &= \int \mathcal{N}(y - x|0, \mathbf{\Sigma}')\mathcal{N}(x|\mu, \mathbf{\Sigma})dx \end{aligned}$$

This is simply the convolution of two multivariate normals, with means 0 and μ and covariances $\mathbf{\Sigma}'$ and $\mathbf{\Sigma}$. The result is also a multivariate normal. We can compute the mean and covariance using Adam's law and the law of total covariance:

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}(\mathbb{E}(Y|X)) \text{ (Adam's Law)} \\ \mathbb{E}(Y|X) &= X \\ \therefore \mathbb{E}(Y) &= \mathbb{E}(X) = \mu. \\ \text{cov}(Y) &= \mathbb{E}(\text{cov}(Y|X)) + \text{cov}(\mathbb{E}(Y|X)) \text{ (LOTC)} \\ &= \mathbb{E}(\mathbf{\Sigma}') + \text{cov}(X) \\ &= \mathbf{\Sigma}' + \mathbf{\Sigma} \\ p(y) &= \mathcal{N}(y|\mu, \mathbf{\Sigma} + \mathbf{\Sigma}') \\ &\boxed{Y \sim \mathcal{N}(\mu, \mathbf{\Sigma} + \mathbf{\Sigma}')} \end{aligned}$$

2. The joint distribution of (X, Y) is also multivariate normal, as proved in §2.3.3 of Bishop. Given that $Y \sim \mathcal{N}(X, \mathbf{\Sigma} + \mathbf{\Sigma}')$, and knowing that the sum of two multivariate normals is also a MVN with the means and covariances summed, we can use the transform $Y = X + Z$, with $Z = \mathcal{N}(0, \mathbf{\Sigma}')$. Therefore, we can find the joint distribution:

$$\begin{aligned} (X, Y) &= (X, X + Z) \\ &\sim \mathcal{N}\left(\begin{bmatrix} \mathbb{E}(X) \\ \mathbb{E}(X + Z) \end{bmatrix}, \begin{bmatrix} \text{var}(X) & \text{cov}(X, X + Z) \\ \text{cov}(X, X + Z) & \text{var}(X + Z) \end{bmatrix} \right) \end{aligned}$$

We know $\mathbb{E}(X) = \mathbb{E}(X + Z) = \mu$, $\text{var}(X) = \mathbf{\Sigma}$, $\text{var}(X + Z) = \mathbf{\Sigma} + \mathbf{\Sigma}'$. Finally, we use linearity of covariance:

$$\text{cov}(X, X + Z) = \text{cov}(X, X) + \text{cov}(X, Z)$$

Since X and Z are independent random variables, $\text{cov}(X, Z) = 0$, and $\text{cov}(X, X) \equiv \text{var}(X)$:

$$\text{cov}(X, X + Z) = \text{var}(X) = \mathbf{\Sigma}$$

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

We confirm that these results are the same as those shown in §2.3.3 of Bishop, where $\mathbf{A} = \mathcal{I}$ and $\mathbf{b} = \mathbf{0}$.

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.

I can Ei-gen – solutions

(a) The non-zero eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ are the same.

Proof: Take $\mathbf{v}$ to be an eigenvector of $\mathbf{X}\mathbf{X}^T$ with eigenvalue λ. Then,

$$\begin{aligned} \mathbf{X}\mathbf{X}^T\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{v} &= \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{v}) = \lambda(\mathbf{X}^T\mathbf{v}) \\ \therefore \mathbf{X}^T\mathbf{X}\mathbf{u} &= \lambda\mathbf{u} \end{aligned}$$

for $\mathbf{u} = \mathbf{X}^T\mathbf{v}$. Therefore, any non-zero eigenvalue of $\mathbf{X}\mathbf{X}^T$ (for eigenvector $\mathbf{v}$) is also an eigenvalue of $\mathbf{X}^T\mathbf{X}$ (for eigenvector $\mathbf{X}^T\mathbf{v}$).

(b) From the spectral theorem, the eigenvalues of a symmetric matrix $\mathbf{X}$ must be real. The eigenvalues of $\mathbf{X} + \mathbf{I}$ are each one greater than the eigenvalues of $\mathbf{X}$.

Proof: Take an eigenvector $\mathbf{v}$ of $\mathbf{X}$ with eigenvalue λ:

$$\begin{aligned} \mathbf{X}\mathbf{v} &= \lambda\mathbf{v} \\ (\mathbf{X} + \mathbf{I})\mathbf{v} &= \mathbf{X}\mathbf{v} + \mathbf{I}\mathbf{v} \\ &= (\lambda + 1)\mathbf{v} \end{aligned}$$

(c) The eigenvalues of $\mathbf{X}^{-1}$ are equal to the reciprocals of the eigenvalues of $\mathbf{X}$.

Proof: Take $\mathbf{v}$ to be an eigenvector of $\mathbf{X}$ with eigenvalue λ, and consider $\mathbf{X}^{-1}\mathbf{v}$:

$$\begin{aligned} \lambda\mathbf{X}^{-1}\mathbf{v} &= \mathbf{X}^{-1}(\lambda\mathbf{v}) \\ &= \mathbf{X}^{-1}(\mathbf{X}\mathbf{v}) \\ &= \mathbf{v} \\ \therefore \mathbf{X}^{-1}\mathbf{v} &= \frac{1}{\lambda}\mathbf{v} \end{aligned}$$

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

Vector Calculus – Solutions

(a)

$$\mathbf{x}^T\mathbf{y} \equiv \sum_i x_i y_i$$

$$\frac{d}{d\mathbf{x}}c \equiv \begin{bmatrix} \frac{d}{dx_0}c \\ \frac{d}{dx_1}c \\ \vdots \end{bmatrix}$$

$$\frac{d}{d\mathbf{x}}\mathbf{x}^T\mathbf{y} = \begin{bmatrix} \frac{d}{dx_0}\sum_i x_i y_i \\ \frac{d}{dx_1}\sum_i x_i y_i \\ \vdots \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \end{bmatrix}$$

$$\boxed{\frac{d}{d\mathbf{x}}\mathbf{x}^T\mathbf{y} = \mathbf{y}}$$

(b)

$$\mathbf{x}^T\mathbf{x} \equiv \sum_i x_i^2$$

$$\frac{d}{d\mathbf{x}}\mathbf{x}^T\mathbf{x} = \begin{bmatrix} \frac{d}{dx_0}\sum_i x_i^2 \\ \frac{d}{dx_1}\sum_i x_i^2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 2x_0 \\ 2x_1 \\ \vdots \end{bmatrix}$$

$$\boxed{\frac{d}{d\mathbf{x}}\mathbf{x}^T\mathbf{x} = 2\mathbf{x}}$$

(c)

$$
\begin{aligned}
(\mathbf{Ax})_i &= \sum_j A_{ij} x_j \\
\mathbf{x}^T \mathbf{Ax} &= \sum_i x_i (\mathbf{Ax})_i = \sum_i \sum_j x_i A_{ij} x_j \\
\left(\frac{d}{d\mathbf{x}} \mathbf{x}^T \mathbf{Ax} \right)_0 &= \frac{d}{dx_0} \sum_i \sum_j x_i A_{ij} x_j \\
&= \sum_j \frac{d}{dx_0} (x_0 A_{0j} x_j) + \sum_i \frac{d}{dx_0} (x_j A_{i0} x_0) \\
&= \sum_j A_{0j} x_j + \sum_i A_{i0} x_i = (\mathbf{Ax} + \mathbf{A}^T \mathbf{x})_0
\end{aligned}
$$

$$
\boxed{\frac{d}{d\mathbf{x}} \mathbf{x}^T \mathbf{Ax} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}}
$$

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

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

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):
   ```

Gradient Check – Solutions

1. My code for the first function:

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

2. My code for the second function:

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

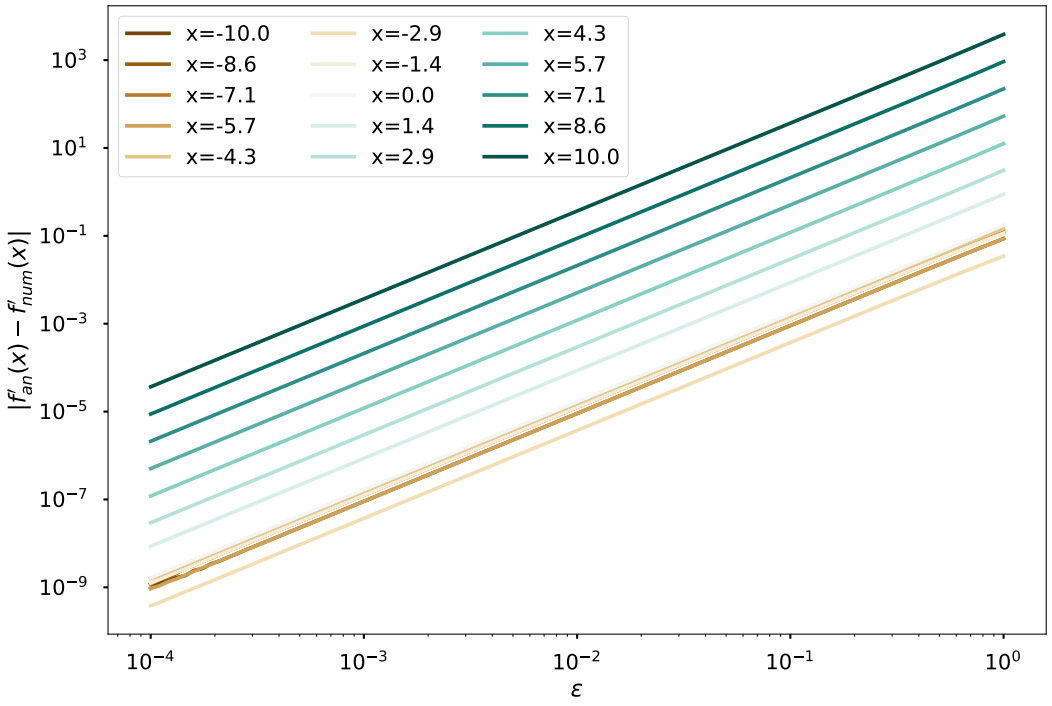
3. My code for the third function:

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

My code for evaluating the comparisons and creating the plot:

```
import matplotlib.pyplot as plt, seaborn.apionly as sns
import matplotlib as mpl
def delta_grad(x, epsilon):
    grad1 = grad_f(x)
    grad2 = grad_check(x, epsilon)
    return np.abs(grad1-grad2)

xs = np.linspace(-10, 10, 15)
epsilons = np.logspace(-4, 0, 100)
results = {}
for x in xs:
    results[x] = delta_grad(x, epsilons)
sns.set_palette(sns.color_palette(''BrBG'', len(xs)))
for k in sorted(results.keys()):
    plt.plot(epsilons, results[k], label='x=%.1f'%k)
plt.legend(loc=0, ncol=3)
plt.xscale('log'), plt.yscale('log')
plt.xlabel(r'$\epsilon$')
plt.ylabel(r'$\left|f^\prime_{an}(x)-f^\prime_{num}(x)\right|$')
plt.savefig('grad_check.pdf')
```



Plot for problem 6 – The magnitude of the disagreement between the analytical ($f'_{an}(x)$) and numerical ($f'_{num}(x)$) gradients are plotted against ϵ for a variety of x. These results show that as ϵ decreases the two results converge.