Hikari Sorensen
laurensorensen@college
CS281-F17

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

Collaborators:

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:

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

Solution

(a) *Proof:*

$$\begin{aligned} cov(X, Y) &= \mathbb{E}((X - \mathbb{E}(X)(Y - \mathbb{E}(Y)) \\ &= \mathbb{E}(XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) \quad \text{since } \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \text{ when } X, Y \text{ are independent} \\ &= 0 \end{aligned}$$

□

(b) Let $X = Z, Y = Z^2$, and $\mathbb{E}(Z) = 0 = \mathbb{E}(Z^3)$. Then

$$\begin{aligned} cov(X, Y) &= \mathbb{E}(z \cdot z^2) - \mathbb{E}(z)\mathbb{E}(z^2) \\ &= \mathbb{E}(Z^3) - \mathbb{E}(z)\mathbb{E}(z^2) \\ &= 0 \end{aligned}$$

but X, Y are not at all independent– indeed, they're highly dependent. □

(c) *Proof:*

$$\begin{aligned} \mathbb{E}(X + aY) &= \sum_x \sum_y (x + ay)P(X = x, Y = y) \\ &= \sum_x \sum_y xP(X = x, Y = y) + \sum_x \sum_y ayP(X = x, Y = y) \\ &= \sum_x \sum_y xP(X = x, Y = y) + \sum_y \sum_x ayP(X = x, Y = y) \\ &\qquad \text{just switching the order of sums in the second term} \\ &= \sum_x x \sum_y P(X = x, Y = y) + a\sum_y y \sum_x P(X = x, Y = y) \\ &\qquad \text{to reflect correct variables being summed over, and since } a \text{ is a constant} \\ &= \sum_x xP(X = x) + a\sum_y yP(Y = y) \\ &= \mathbb{E}(X) + a\mathbb{E}(Y) \end{aligned}$$

□

$$
\begin{aligned}
\mathrm{var}(X + aY) &= \mathbb{E}(((X + aY) - \mathbb{E}(X + aY))^2) \\
&= \mathbb{E}(((X + aY) - (\mathbb{E}(X) + a\mathbb{E}(Y)))^2) \quad \text{by linearity of expectations (above)} \\
&= \mathbb{E}(((X - \mathbb{E}(X)) + a(Y - \mathbb{E}(Y)))^2) \\
&= \mathbb{E}((X - \mathbb{E}X)^2 + 2a(X - \mathbb{E}X)(Y - \mathbb{E}Y) + a^2(Y - \mathbb{E}Y)^2) \\
&\quad \text{where I've at this point switched notation such that } \mathbb{E}Z = \mathbb{E}(Z), \text{ since the brackets} \\
&\quad \text{are getting confusingly messy} \\
&= \mathbb{E}(X - \mathbb{E}X)^2 + 2a\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) + a^2\mathbb{E}(Y - \mathbb{E}Y)^2 \quad \text{by linearity} \\
&= \mathbb{E}(X - \mathbb{E}X)^2 + a^2\mathbb{E}(Y - \mathbb{E}Y)^2 \\
&\quad \text{since the middle term is the covariance, which is } 0 \text{ for } X, Y \text{ independent} \\
&= \mathrm{var}(X) + a^2\mathrm{var}(Y)
\end{aligned}
$$

□

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.

Solution

(a) Yes, as long as it still integrates to 1. For example, take the uniform distribution on the interval $[0, a]$ for $a < 1$. Then in order for the total probability to equal 1, we must have that the pdf $f(x) = \frac{1}{a} > 1$ over the interval $[0, a]$ (and $f(x) = 0$ otherwise).

(b) Let $f(x)$ be the pdf of X. Then

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

(c)

$$f(0) = \sqrt{\frac{50}{x}}$$

(d)

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

(e) In short, a probability *density* is not the same as a probability. Since a PDF is defined over a continuous (in particular, uncountably infinite) interval, the probability that it will take on any particular value is 0 (consider that if not, then we would sum an uncountably infinite number of positive values, which would certainly not have sum equal to 1, as is required of a probability distribution). Only intervals of positive length have nonzero probability. This is why in part (a), we have that a pdf can take values greater than 1, despite probabilities themselves being unable to have values greater than 1.

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

(a) If $Y|X = x$ has mean x, then we can interpret Y as being some observation of X plus some normally distributed noise around x, but independent of X. Thus, let $Y = X + Z$, where $Z \sim \mathcal{N}(0, \Sigma')$. Then

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

by the closure under addition of means and variances of independent MVN's.
As a check of our guess, if we consider $Y|X$ again, we have that

$$Y|X = x \sim x + Z \sim \mathcal{N}(x, \Sigma'),$$

as desired.

(b)

$$\begin{aligned} P(X, Y) &= P(X)P(Y|X) \\ &= |2\pi\mathbf{\Sigma}|^{-1/2} \exp(-\frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma^{-1}} \mathbf{x}) \cdot |2\pi\mathbf{\Sigma}'|^{-1/2} \exp(-\frac{1}{2}(\mathbf{y}-\mathbf{x})^\top \mathbf{\Sigma'^{-1}}(\mathbf{y}-\mathbf{x})) \\ &= (2\pi)^{-1}|\mathbf{\Sigma\Sigma}'|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}^\top \mathbf{\Sigma^{-1}} \mathbf{x} + (\mathbf{y}-\mathbf{x})^\top \mathbf{\Sigma'^{-1}}(\mathbf{y}-\mathbf{x}))) \\ &= \text{ more algebra that I can't get through before my time's up...} \end{aligned}$$

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) Let v be an eigenvector of $\mathbf{X}\mathbf{X}^\top$, with eigenvalue $\lambda \neq 0$. Then

$$\begin{aligned} & \mathbf{X}\mathbf{X}^\top \mathbf{v} = \lambda \mathbf{v} \\ \implies & \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top \mathbf{v}) = \mathbf{X}^\top(\lambda \mathbf{v}) \\ \implies & (\mathbf{X}^\top \mathbf{X})(\mathbf{X}^\top \mathbf{v}) = \lambda(\mathbf{X}^\top \mathbf{v}), \end{aligned}$$

and so $\mathbf{X}^\top \mathbf{v}$ is an eigenvector of $X^\top X$ with eigenvalue λ.
Thus, $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$ have the same nonzero eigenvalues.

(b) Let $\mathbf{v} \neq \mathbf{w}$ be eigenvectors of X, corresponding to nonzero eigenvalues $\lambda_1 \neq \lambda_2$, respectively. Then

$$\begin{aligned} & \lambda_1 \mathbf{v} = \mathbf{X}\mathbf{v} \\ \implies & \mathbf{w}^\top \lambda_1 \mathbf{v} = \mathbf{w}^\top \mathbf{X}\mathbf{v} \\ \implies & \lambda_1 \mathbf{w}^\top \mathbf{v} = (\mathbf{X}^\top \mathbf{w})^\top \mathbf{v} \\ & \qquad = (\mathbf{X}\mathbf{w})^\top \mathbf{v} \qquad \text{since } \mathbf{X} \text{ is symmetric} \\ & \qquad = (\lambda_2 \mathbf{w})^\top \mathbf{v} \\ & \qquad = \lambda_2 \mathbf{w}^\top \mathbf{v} \\ \implies & (\lambda_1 - \lambda_2)\mathbf{w}^\top \mathbf{v} = 0 \end{aligned}$$

But now since we said that $\lambda_1 \neq \lambda_2$, it must be that $\mathbf{w}^\top \mathbf{v} = 0$, and thus $\mathbf{w}, \mathbf{v}$ are orthogonal.
Thus, if the eigenvalues of a set of eigenvectors of $\mathbf{X}$ are distinct, then those eigenvectors are mutually orthogonal.

(c) If λ is an eigenvalue of $\mathbf{X}$, then $\frac{1}{\lambda}$ is an eigenvalue of $\mathbf{X}^{-1}$. That is, the eigenvalues of $\mathbf{X}^{-1}$ are the inverses of the eigenvalues of $\mathbf{X}$.

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) Let $f = \mathbf{x}^\top \mathbf{y} = x_1 y_1 + ... + x_m y_m$. Then

$$\begin{aligned}\nabla_{\mathbf{x}} f &= (\partial_{x_1} f, ..., \partial_{x_m} f) \\ &= (y_1, ..., y_m) \\ &= \mathbf{y}\end{aligned}$$

(b) Let $g = \mathbf{x}^\top \mathbf{x} = x_1 x_1 + ... + x_m x_m = x_1^2 + ... + x_m^2$. Then

$$\begin{aligned}\nabla_{\mathbf{x}} g &= (\partial_{x_1} g, ..., \partial_{x_m} g) \\ &= (2x_1, ..., 2x_m) \\ &= 2 \cdot (x_1, ..., x_m) \\ &= 2\mathbf{x}\end{aligned}$$

(c) Let $h = \mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^m \sum_{j=1}^m x_i A_{ij} x_j$. Then

$$\begin{aligned}\nabla_{\mathbf{x}} h &= (\partial_{x_1} h, ..., \partial_{x_m} h) \\ &= (\sum_{i=1}^m A_{i1} x_i + \sum_{j=1}^m A_{1j} x_j, ..., \sum_{i=1}^m A_{im} x_i + \sum_{j=1}^m A_{mj} x_j) \\ &= (\sum_{i=1}^m (A_{i1} + A_{1i}) x_i, ..., \sum_{i=1}^m (A_{im} + A_{mi})) x_i \qquad \text{just by reindexing } j\text{'s with } i\text{'s} \\ &= (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}\end{aligned}$$

And thus if $\mathbf{A}$ is symmetric,

$$\nabla_{\mathbf{x}} h = 2\mathbf{A}\mathbf{x}.$$

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):
   ```

Solution

```
import numpy as np
```

1.

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

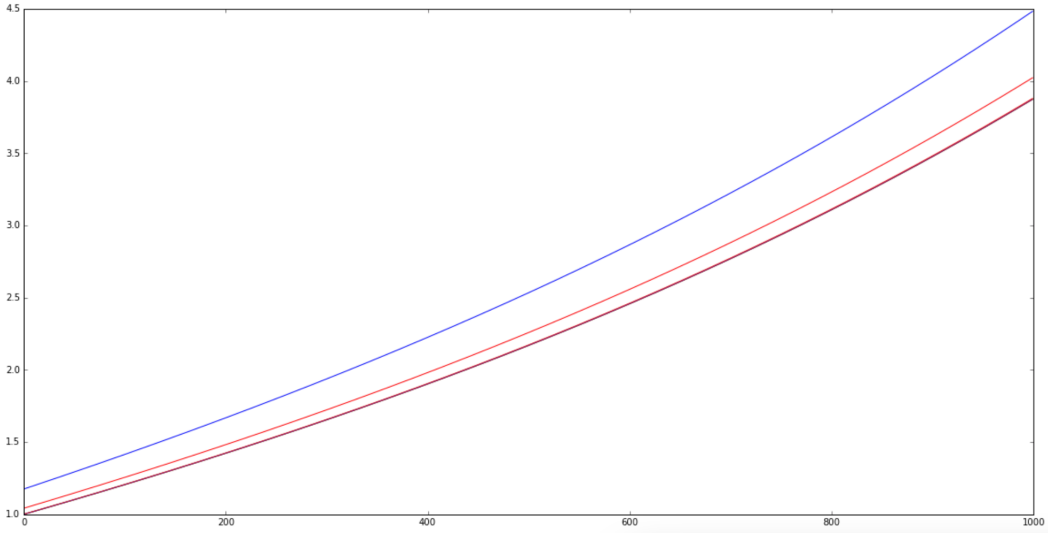
2.

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

3.

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

The plot below shows the values of the analytical gradient of $f(x)$ as compared with the numerical approximation, for $x \in [0, 1]$ at 0.001 increments.



The blue line is the approximation at $\epsilon = 1.0$.
The red line is the approximation at $\epsilon = 0.5$.
The purple-ish line is actually four lines: the approximations at $\epsilon = 0.1$, at $\epsilon = 0.01$, and at $\epsilon = 0.001$, as well as the analytically computed gradient. However, as can be seen clearly, the approximations converse sufficiently quickly that we cannot visually discern between the approximations and the true gradient.