Anirudh Suresh
anirudh_suresh@college.harvard.edu
CS281-F17

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

Collaborators: John Doe, Fred Doe

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) Using the standard definition of covariance,

$$Cov(X, Y) = \mathbb{E}(XY) - \mu_X\mu_Y.$$

When X and Y are independent, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Therefore,

$$Cov(X, Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mu_X\mu_Y = \mu_X\mu_Y - \mu_X\mu_Y = 0.$$

Hence, the covariance of two independent random variables is 0.

(b) Let us use X and $Y = X^2$ as our two random variables. Clearly, Y is not independent of X; its very value depends completely on the value of X. However, if we look at the covariance calculation,

$$Cov(X, Y) = Cov(X, X^2) = \mathbb{E}(XX^2) - \mu_X\mu_{X^2} = \mathbb{E}(X^3) - 0(\mu_{X^2}) = 0.$$

Because X and X^3 are symmetric about $X = 0$, their expectations are 0 for any range symmetric about the y-axis. Hence, though X and Y are inherently tied together in this example, they have zero covariance, indicating that zero covariance does not imply independence.

(c) Because of the independence of X and Y,

$$\mathbb{E}(X + aY) = \mathbb{E}(X) + \mathbb{E}(aY).$$

Using the standard definition of covariance and realizing that since a is a constant, $Cov(a, Y) = 0$,

$$Cov(a, Y) = \mathbb{E}(aY) - a\mu_Y = 0$$
$$\mathbb{E}(aY) = a\mu_Y = a\mathbb{E}(Y).$$

Hence,

$$\mathbb{E}(X + aY) = \mathbb{E}(X) + \mathbb{E}(aY) = \mathbb{E}(X) + a\mathbb{E}(Y).$$

For the second property, we can use a definition of variance in terms of covariance.

$$\text{var}(X + aY) = \text{var}(X) + 2Cov(X, aY) + \text{var}(aY) = \text{var}(X) + 0 + \text{var}(aY)$$
$$\text{var}(aY) = \mathbb{E}(a^2Y^2) - (\mathbb{E}(aY))^2 = a^2\mathbb{E}(Y^2) - (a\mathbb{E}(Y))^2 = a^2(\mathbb{E}(Y^2) - (\mathbb{E}(Y))^2) = a^2\text{var}(Y).$$

Thus,

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

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) A PDF can certainly take values greater than 1. For example, any random variable uniformly distributed over a range smaller than 1 has a PDF that takes on values greater than 1 in that range.

(b) $f(x) = \frac{1}{\sqrt{2\pi(\frac{1}{100})}} \exp(-\frac{x^2}{2(\frac{1}{100})}) = \frac{10}{\sqrt{2\pi}} \exp(-50x^2)$

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

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

(e) This is a continuous random variable, and thus, the probability that it is equal to any one discrete value is infinitesimally small. Hence, although $X = 0$ is the obvious center of the distribution, it by itself has 0 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

(a) Using Adam's Law,

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

Using Eve's Law,

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

$Y \mid X \sim \mathcal{N}$, and through the closure properties of the multivariate normal (if the conditional on a Gaussian is Gaussian, the marginal should similarly be Gaussian), $Y \sim \mathcal{N}$. Thus,

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

(b) We know that the joint distribution for the pair (X, Y) given that both variables have a marginal normal distribution and the conditional is normal is

$$(X, Y) \sim \mathcal{N}(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \text{var}(X) & Cov(X, Y) \\ Cov(Y, X) & \text{var}(Y) \end{pmatrix}).$$

We can compute the covariance from the definition of covariance

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Via Adam's Law,

$$= \mathbb{E}(\mathbb{E}(XY \mid X)) - (\mu)(\mu)$$
$$= \mathbb{E}(X^2) - \mu^2$$
$$\mathbb{E}(X^2) = \text{var}(X) + (\mathbb{E}(X))^2 = \Sigma + \mu^2$$

Thus,

$$Cov(X, Y) = \Sigma + \mu^2 - \mu^2 = \Sigma.$$

Hence,

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

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.

Solution

(a) Using the definition of an eigenvalue-eigenvector pair,

$$\mathbf{X}^T\mathbf{X}\mathbf{v_1} = \lambda\mathbf{v_1}.$$

Left multiplying each side of the identity above by $\mathbf{X}$,

$$\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{v_1} = \lambda\mathbf{X}\mathbf{v_1}$$

$$\mathbf{X}\mathbf{X}^T\mathbf{v_2} = \lambda\mathbf{v_2},$$

where $\mathbf{v_2} = \mathbf{X}\mathbf{v_1}$ reveals that $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ share the same eigenvalues.

(b) If $\mathbf{X}$ is square and symmetric, it has the same number of (not necessarily distinct) eigenvalues as its dimensions/rank. Hence, it will have n real (if all its entries are real) eigenvalues. In addition,

$$\mathbf{X} = \mathbf{X}^T.$$

Thus, the eigenvalues of $\mathbf{X}$ are the same as the eigenvalues of $\mathbf{X}^T$. If we denote the eigenvalues of $\mathbf{X}$ as λ_i, $i \in [1, n]$ and the eigenvalues of $\mathbf{X} + \mathbf{I}$ as κ_i, $i \in [1, n]$,

$$\kappa_i = \lambda_i + 1, i \in [1, n].$$

(c) Via the spectral theorem,

$$\mathbf{X} = \mathbf{Q}^{-1}\mathbf{D}\mathbf{Q},$$

where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{D}$ is a diagonalized form of $\mathbf{X}$. $\mathbf{X}^{-1}$. Then, we can immediately see that

$$\mathbf{X}^{-1} = \mathbf{Q}^{-1}\mathbf{D}^{-1}\mathbf{Q}.$$

Since $\mathbf{D}$ is a diagonal matrix, $\mathbf{D}^{-1}$ is as well, with its diagonal entries the reciprocals of their respective counterparts in $\mathbf{D}$. Since the diagonal matrix $\mathbf{D}$ contains the eigenvalues of $\mathbf{X}$. $\mathbf{D}^{-1}$ must contain the eigenvalues of $\mathbf{X}^{-1}$. Hence, given eigenvalues of $\mathbf{X}$ $\lambda_i, i \in [1, n]$, $\mathbf{X}^{-1}$ must have eigenvalues $\frac{1}{\lambda_i}, i \in [1, n]$. For this to be the case, $\lambda_i \neq 0$ for all $i \in [1, n]$, which is the case since $\mathbf{X}$ is specified to be invertible.

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) Using the symbol $:=:$ to indicate the scalar "equivalent" of a particular vector notation term,

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{m} x_i y_i :=: xy$$

$$\frac{\partial \mathbf{x}^T\mathbf{y}}{\partial \mathbf{x}} = \frac{\partial(\sum_{i=1}^{m} x_i y_i)}{\partial \mathbf{x}} :=: \frac{\partial(xy)}{\partial x}$$

$$\frac{\partial \mathbf{x}^T\mathbf{y}}{\partial \mathbf{x}} = \mathbf{y}$$

(b)

$$\mathbf{x}^T\mathbf{x} = \sum_{i=1}^{m} x_i x_i :=: xx = x^2$$

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

$$\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$$

(c) Using results from the previous two parts and the product rule,

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial(\sum_{i=1}^{m} x_i(\sum_{j=1}^{m}\sum_{k=1}^{m} A_{j,k}x_k)_i)}{\partial \mathbf{x}} = \frac{\partial(\sum_{j=1}^{m}\sum_{k=1}^{m} A_{j,k}x_k x_j)}{\partial \mathbf{x}}$$

$$\partial(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\partial\mathbf{x})^T\mathbf{A}\mathbf{x} + \mathbf{x}\mathbf{A}\partial\mathbf{x} = (\partial\mathbf{x})^T\mathbf{A}\mathbf{x} + (\partial\mathbf{x})^T\mathbf{A}^T\mathbf{x}$$
$$= (\partial\mathbf{x})^T(\mathbf{A} + \mathbf{A}^T)\mathbf{x} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)\partial\mathbf{x}$$

Therefore,

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T).$$

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):
   ```

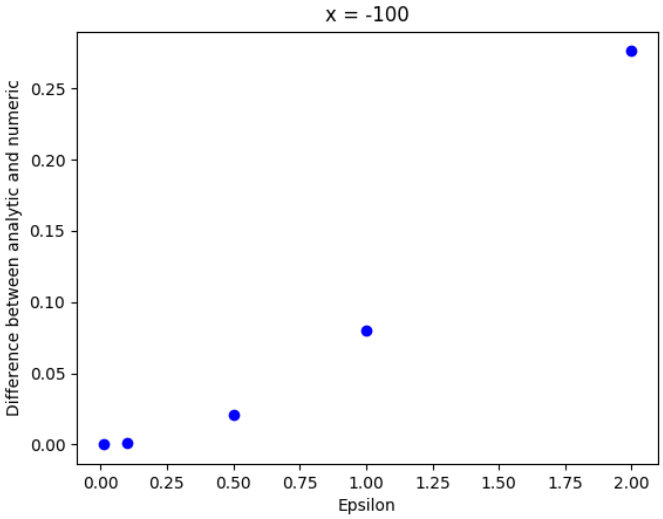
3. Now, implement a gradient check (the numerical approximation to the derivative), and by plotting, show that the numerical approximation approaches the analytic as `epsilon` $\to 0$ for a few values of x:

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

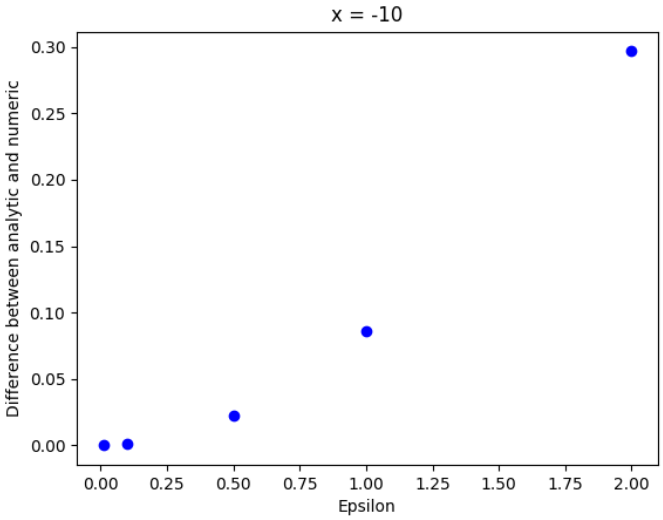
Solution

1. See Python script
2. $$f'(x) = -\sin(x) + 2x + e^x$$

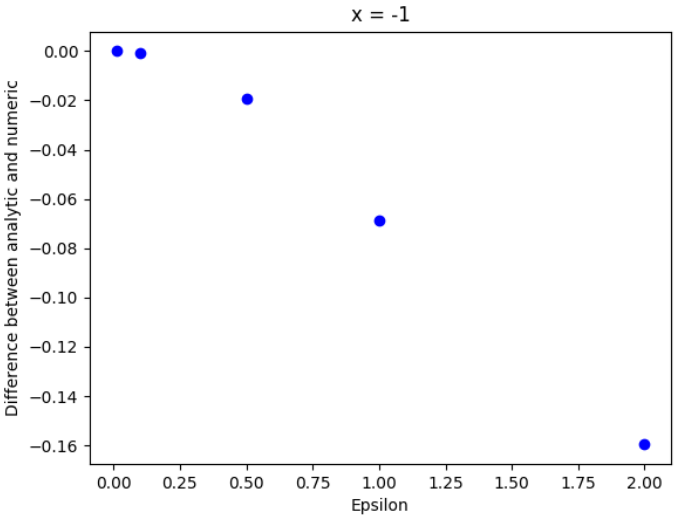
   See Python script for implementation
3. See Python script for implementation; the plots for 7 different values of x below clearly reveal that as $\epsilon \to 0$, the difference between the analytic and numerical expressions for the gradient converges to 0.



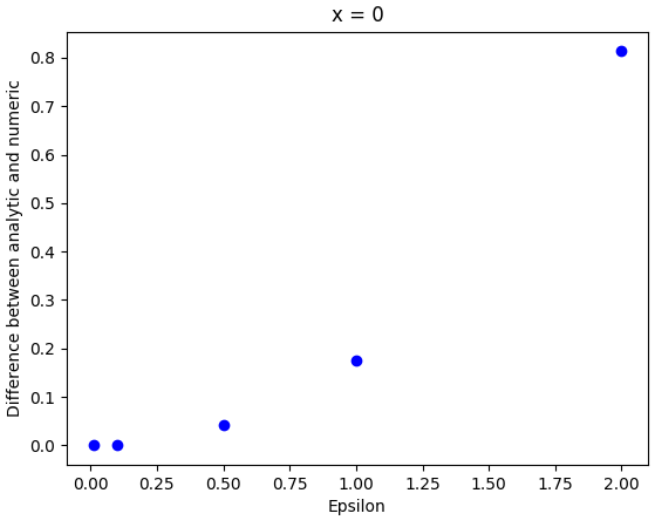
Gradient check for x = -100



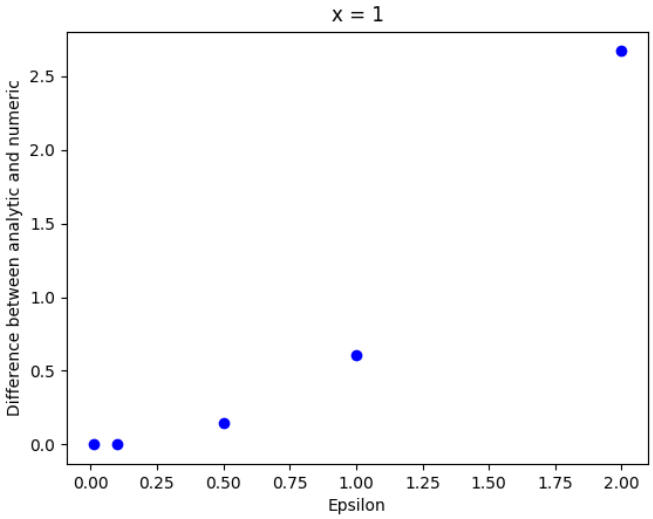
Gradient check for x = -10



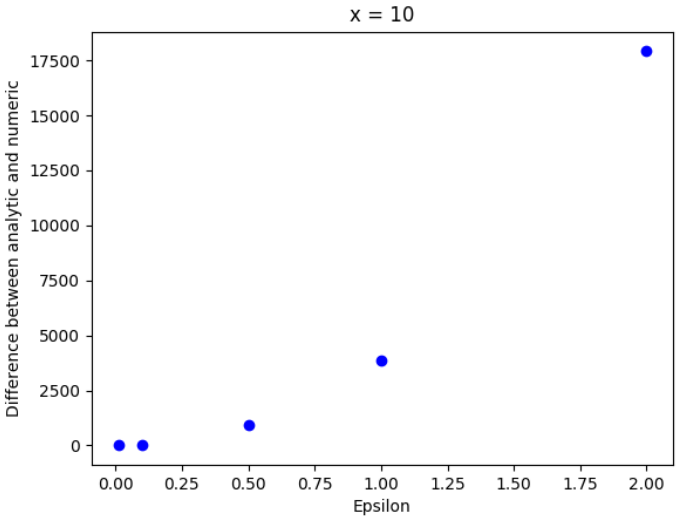
Gradient check for x = -1



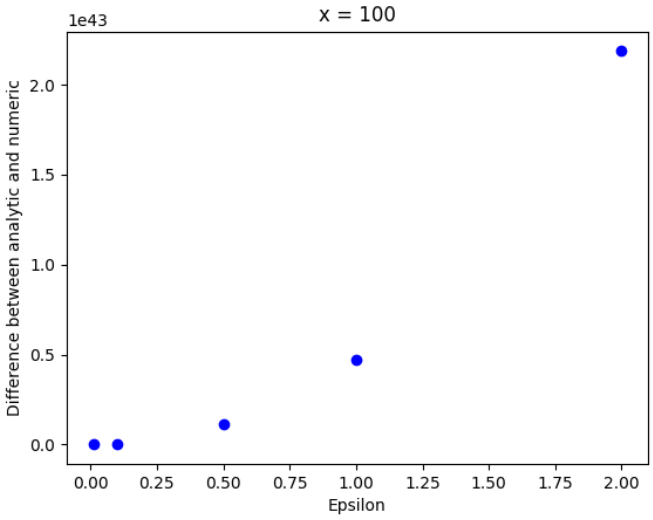
Gradient check for x = 0



Gradient check for x = 1



Gradient check for x = 10



Gradient check for x = 100