Christopher Hase
christopher_hase@g.harvard.edu
CS281-F17

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

Collaborators: None

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:

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

(a) First we'll show that $X \perp\!\!\!\perp Y \Rightarrow E(XY) = E(X)E(Y)$, assuming X and Y are discrete random variables.

Using the definition of expectation, we get that:

$E(XY) = \sum_x \sum_y xyP(X = x, Y = y)$

$= \sum_x \sum_y xyP(X = x)P(Y = y)$

$= \sum_x xP(X = x) \sum_y yP(Y = y)$

$= E(X)E(Y)$

The same can be shown for independent, continuous random variables. Now we can show (a) easily:

$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$

(b) We'll use the properties of expectation, variance, covariance, and Bernoulli random variables in the following example.

Let $X \sim Bern\,(p)\,, -Y \sim Bern\,(p)$

$Cov\,(X + Y, X - Y) = Cov\,(X, X) - Cov\,(X, Y) + Cov\,(Y, X) - Cov\,(Y, Y)$

$= Cov\,(X, X) - Cov\,(Y, Y)$

$= Var\,(X) - Var\,(Y) = 0$

$E\,(X - Y) = E\,(X) + E\,(-Y) = 2p \neq E\,(X - Y|X + Y = 1) = 1$

Thus, we have $Cov(X + Y, X - Y) = 0$ but $X + Y$ and $X - Y$ not independent since

$E(X - Y) \neq E(X - Y|X + Y = 1)$

(c) For proof the first property, we'll assume X and Y are discrete random variables and use the definition of expectation as well as properties of independent random variables.

$E(X) + aE(Y) = \sum_x xP(X = x) + a\sum_y yP(Y = y)$

$= \sum_x xP(X = x)\sum_y P(Y = y) + \sum_y ayP(Y = y)\sum_x P(X = x)$

$= \sum_x \sum_y xP(X = x)P(Y = y) + \sum_x \sum_y ayP(Y = y)P(X = x)$

$= \sum_x \sum_y (x + ay)P(X = x)P(Y = y)$

$= \sum_x \sum_y (x + ay)P(X = x, Y = y) = E(X + aY)$

For proof of the second property, we'll use properties of independent random variables, expectation, and variance.

$Var(X + aY) = E\left[(X + aY)^2\right] - [E(X + aY)]^2$

$= E\left[X^2 + 2aXY + a^2Y^2\right] - [E(X) + aE(Y)]^2$

$= E\left(X^2\right) + 2aE(XY) + a^2E\left(Y^2\right) - [E(X)]^2 - 2aE(X)E(Y) - a^2[E(Y)]^2$

$= Var(X) + a^2\left(E\left(Y^2\right) - [E(Y)]^2\right)$

$= Var(X) + a^2Var(Y)$

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.

(a) Yes. For example, let $X \sim Unif\left(0, \frac{1}{10}\right)$. Then $p(x) = 10 \; \forall x \in \left[0, \frac{1}{10}\right]$. For continuous random variables, the integral over the support needs to be 1. For discrete random variables, the sum over the support needs to be 1.

(b) $X \sim N\left(0, \frac{1}{100}\right) \Rightarrow p(x) = \frac{10}{\sqrt{2\pi}} \exp\left(-50x^2\right)$

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

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

(e) The value of the PDF for a random variable X is not the probability that X takes on a particular value x. The value of the PDF at some x that X takes on is the height of the distribution at x. The probability of some event associated with the value of X is the area under the PDF that represents that event. The area under the PDF that represents any one particular value of x that X takes on is 0. This explains why $p(0) \neq P(X = 0) = 0$.

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) We can see that $\mathbf{Y} - \mathbf{X} | \mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}')$. This distribution is not dependent on $\mathbf{X} \Rightarrow \mathbf{Y} - \mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}')$. Since $(\mathbf{Y} - \mathbf{X}) \perp\!\!\!\perp \mathbf{X}$, the sum of independent multivariate normal random vectors is a multivariate normal random vector, and $\mathbf{Y} - \mathbf{X} + \mathbf{X} = \mathbf{Y}$, it follows that $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma} + \mathbf{\Sigma}')$.

(b) We are trying to find the joint distribution of $(\mathbf{X}, \mathbf{Y})^T$. Since $\mathbf{X}$ is MVN and $\mathbf{Y}|\mathbf{X}$ is MVN, $(\mathbf{X}, \mathbf{Y})^T \sim N(\boldsymbol{\theta}, \boldsymbol{\beta})$ such that $\boldsymbol{\theta} \in \mathbb{R}^{2m}$ and $\boldsymbol{\beta} \in \mathbb{R}^{2m \times 2m}$.

According to Theorem 4.3.1 in Murphy's text, $\boldsymbol{\theta} = (E(\mathbf{X}), E(\mathbf{Y}))^T = (\boldsymbol{\mu}, \boldsymbol{\mu})^T$

Also according to Theorem 4.3.1 in Murphy's text, $\boldsymbol{\beta} = \begin{pmatrix} Var(\mathbf{X}) & Cov(\mathbf{X}, \mathbf{Y}) \\ Cov(\mathbf{Y}, \mathbf{X}) & Var(\mathbf{Y}) \end{pmatrix}$.

We know that $Var(\mathbf{X}) = \mathbf{\Sigma}$ and that $Var(\mathbf{Y}) = \mathbf{\Sigma} + \mathbf{\Sigma}'$, so all we need to do is find $Cov(\mathbf{X}, \mathbf{Y})$ and $Cov(\mathbf{Y}, \mathbf{X})$.

$Cov(\mathbf{X}, \mathbf{Y}) = E(Cov(\mathbf{X}, \mathbf{Y})|\mathbf{X}) + Cov(E(\mathbf{X}|X), E(\mathbf{Y}|\mathbf{X})) = \mathbf{0} + Cov(\mathbf{X}, \mathbf{X}) = Var(\mathbf{X}) = \mathbf{\Sigma}$

$Cov(\mathbf{Y}, \mathbf{X}) = E(Cov(\mathbf{Y}, \mathbf{X})|\mathbf{X}) + Cov(E(\mathbf{Y}|\mathbf{X}), E(\mathbf{X}|\mathbf{X})) = \mathbf{0} + Cov(\mathbf{X}, \mathbf{X}) = Var(\mathbf{X}) = \mathbf{\Sigma}$

Then, $\boldsymbol{\beta} = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} \\ \mathbf{\Sigma} & \mathbf{\Sigma} + \mathbf{\Sigma}' \end{pmatrix}$.

Therefore, $(\mathbf{X}, \mathbf{Y})^T \sim N((\boldsymbol{\mu}, \boldsymbol{\mu})^T, \begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} \\ \mathbf{\Sigma} & \mathbf{\Sigma} + \mathbf{\Sigma}' \end{pmatrix})$.

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) Suppose λ is a non-zero eigenvalue of $\mathbf{X}^T\mathbf{X}$ and is associated with eigenvector $\mathbf{a} \in \mathbb{R}^m$. Then $\mathbf{X}^T\mathbf{X}\mathbf{a} = \lambda\mathbf{a} \Rightarrow (\mathbf{X}\mathbf{X}^T)\mathbf{X}\mathbf{a} = \lambda\mathbf{X}\mathbf{a} \Rightarrow \mathbf{X}\mathbf{X}^T$ also has λ as an eigenvalue and it is associated with eigenvector $\mathbf{X}\mathbf{a}$. Since the choice of λ was arbitrary, it follows that the non-zero eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ are the same.

It follows that in the case that $n = m$, $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ will have the same eigenvalues.

If $n \neq m$, the larger of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ will have more eigenvalues with value 0. This is because $\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}\mathbf{X}^T) \leq \min(n, m)$, and these ranks correspond to the maximum number of non-zero eigenvalues that $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ can have $\Rightarrow \mathbf{X}^T\mathbf{X}$ has $m - n$ more eigenvalues with value 0 than $\mathbf{X}\mathbf{X}^T$ has if $m > n$ or $n - m$ less eigenvalues with value 0 than $\mathbf{X}\mathbf{X}^T$ has if $m < n$.

(b) The spectral theorem says that if $\mathbf{X}$ is a square, symmetric matrix, all the eigenvalues of $\mathbf{X}$ are real.

Let λ be an eigenvalue of $\mathbf{X}$ associated with eigenvector $\mathbf{a}$. For the eigenvalues of $\mathbf{X} + \mathbf{I}$, we start with $\mathbf{X}\mathbf{a} = \lambda\mathbf{a} \Rightarrow (\mathbf{X} + \mathbf{I})\mathbf{a} = (\lambda\mathbf{I} + \mathbf{I})\mathbf{a} \Rightarrow (\mathbf{X} + \mathbf{I})\mathbf{a} = (\lambda + 1)\mathbf{a}$. Since the choice of λ was arbitrary, it follows the eigenvalues of $\mathbf{X} + \mathbf{I}$ can be found by adding 1 to each of the eigenvalues of $\mathbf{X}$.

(c) Suppose λ is an eigenvalue of $\mathbf{X} \in \mathbb{R}^{n \times n}$ (λ must be non-zero since $\mathbf{X}$ is invertible) and is associated with eigenvector $\mathbf{a} \in \mathbb{R}^n$. Then $\mathbf{X}\mathbf{a} = \lambda\mathbf{a} \Rightarrow \mathbf{X}^{-1}\mathbf{X}\mathbf{a} = \lambda\mathbf{X}^{-1}\mathbf{a} \Rightarrow \mathbf{a} = \lambda\mathbf{X}^{-1}\mathbf{a} \Rightarrow \mathbf{X}^{-1}\mathbf{a} = \frac{1}{\lambda}\mathbf{a} \Rightarrow \mathbf{X}^{-1}$ has $\frac{1}{\lambda}$ as an eigenvalue and it is associated with eigenvector $\mathbf{a}$. Since the choice of λ was arbitrary, it follows the eigenvalues of $\mathbf{X}$ and $\mathbf{X}^{-1}$ are reciprocals of each other.

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)

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}^T\mathbf{y}\right) = \frac{\partial}{\partial \mathbf{x}}\left(\sum_{i=1}^{m} x_i y_i\right) = \sum_{i=1}^{m} \frac{\partial x_i y_i}{\partial \mathbf{x}} = \sum_{i=1}^{m}\left[\frac{\partial x_i y_i}{\partial x_1}, \frac{\partial x_i y_i}{\partial x_2} \cdots \frac{\partial x_i y_i}{\partial x_m}\right]^T = [y_1, y_2 \ldots y_m]^T = \mathbf{y}$$

(b)

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}^T\mathbf{x}\right) = \frac{\partial}{\partial \mathbf{x}}\left(\sum_{i=1}^{m} x_i^2\right) = \sum_{i=1}^{m} \frac{\partial x_i^2}{\partial \mathbf{x}} = \sum_{i=1}^{m}\left[\frac{\partial x_i^2}{\partial x_1}, \frac{\partial x_i^2}{\partial x_2} \cdots \frac{\partial x_i^2}{\partial x_m}\right]^T = [2x_1, 2x_2, \ldots 2x_m]^T = 2\mathbf{x}$$

(c)

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right) = \frac{\partial}{\partial \mathbf{x}}\left(x_1 \sum_{i=1}^{m} x_i a_{i1} + x_2 \sum_{i=1}^{m} x_i a_{i2} + \ldots + x_m \sum_{i=1}^{m} x_i a_{im}\right)$$

$$= \sum_{i=1}^{m} \frac{\partial x_1 x_i a_{i1}}{\partial \mathbf{x}} + \ldots + \sum_{i=1}^{m} \frac{\partial x_m x_i a_{im}}{\partial \mathbf{x}}$$

$$= \sum_{i=1}^{m}\left[\frac{\partial x_1 x_i a_{i1}}{\partial x_1} \cdots \frac{\partial x_1 x_i a_{i1}}{\partial x_m}\right]^T + \ldots + \sum_{i=1}^{m}\left[\frac{\partial x_m x_i a_{im}}{\partial x_1} \ldots \frac{\partial x_m x_i a_{im}}{\partial x_m}\right]^T$$

$$= \left[2x_1 a_{11} + \sum_{i=2}^{m} x_i a_{i1}, \ldots, x_1 a_{m1}\right]^T + \ldots + \left[x_m a_{1m}, \ldots, 2x_m a_{mm} + \sum_{i=1}^{m-1} x_i a_{im}\right]^T = \left(\mathbf{A} + \mathbf{A}^T\right)\mathbf{x}$$

We can see that this is true by doing a manual calculation using a small value for m.

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:

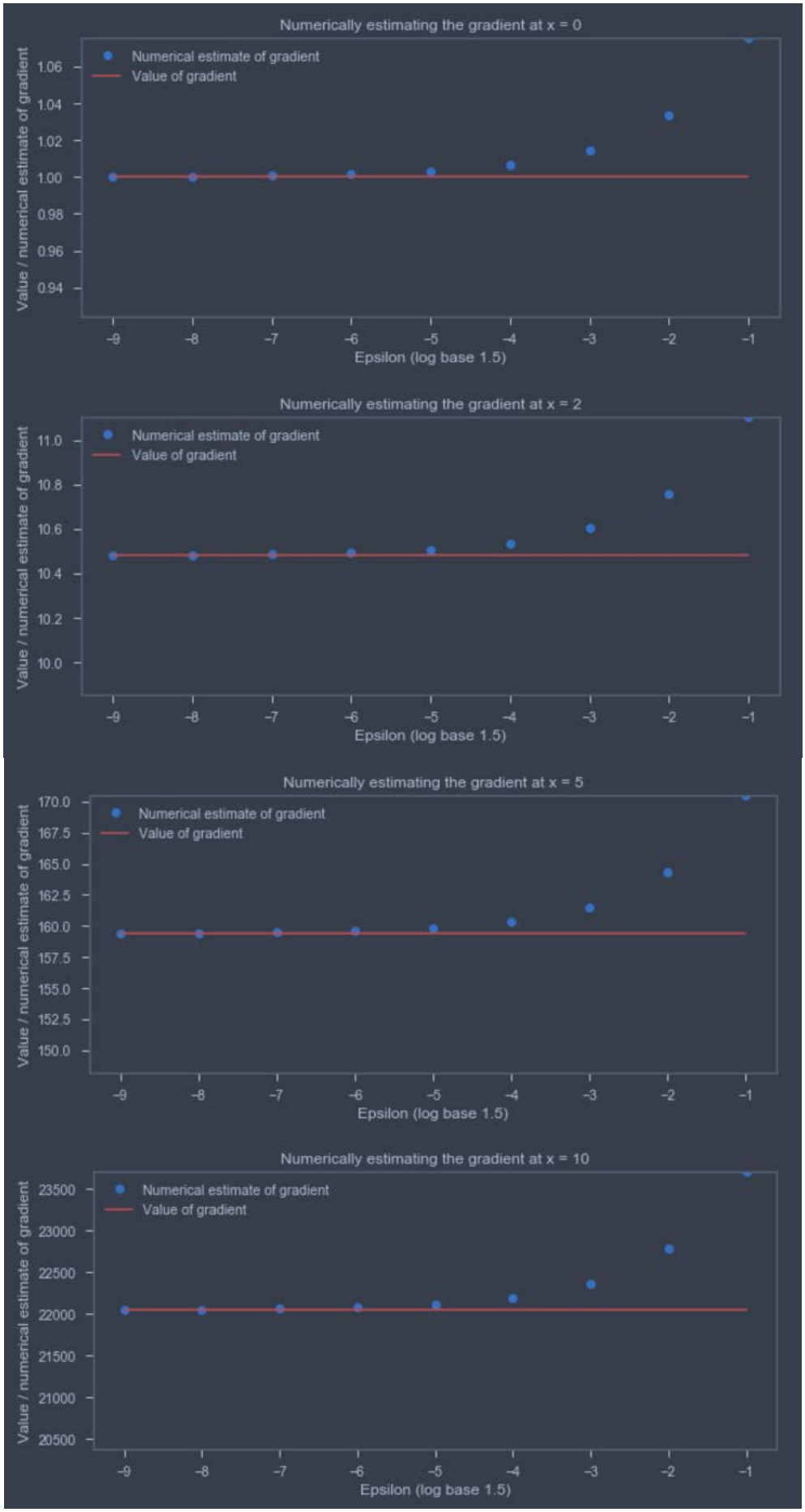
   ```
   def grad_f(x):
   ```

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):
   ```

$$\frac{df}{dx} = -\sin(x) + 2x + e^x$$

The plots on the following page show that the numerical approximations of the gradient approach the analytic values of the gradient as $\epsilon \to 0$ for four values of x.


Numerically estimating the gradient at x = 0
Numerical estimate of gradient
Value of gradient
Value / numerical estimate of gradient
1.06
1.04
1.02
1.00
0.98
0.96
0.94
−9
−8
−7
−6
−5
−4
−3
−2
−1
Epsilon (log base 1.5)
Numerically estimating the gradient at x = 2
Numerical estimate of gradient
Value of gradient
Value / numerical estimate of gradient
11.0
10.8
10.6
10.4
10.2
10.0
−9
−8
−7
−6
−5
−4
−3
−2
−1
Epsilon (log base 1.5)
Numerically estimating the gradient at x = 5
Numerical estimate of gradient
Value of gradient
Value / numerical estimate of gradient
170.0
167.5
165.0
162.5
160.0
157.5
155.0
152.5
150.0
−9
−8
−7
−6
−5
−4
−3
−2
−1
Epsilon (log base 1.5)
Numerically estimating the gradient at x = 10
Numerical estimate of gradient
Value of gradient
Value / numerical estimate of gradient
23500
23000
22500
22000
21500
21000
20500
−9
−8
−7
−6
−5
−4
−3
−2
−1
Epsilon (log base 1.5)