Homework 1

January 26, 2022

1 Basic calculus

1. Calculate the derivative of $f(x)$

   (a) $f(x) = e^x$

   $$f'(x) = e^x$$

   (b) $f(x) = \log(1 + x)$

   $$f'(x) = \frac{1}{1+x}$$

   c) $f(x) = \log(1 + e^x)$

   $$f'(x) = \frac{e^x}{1+x}$$

2. Taylor expansion. Let f: $\mathbb{R} \to \mathbb{R}$ be a twice differentiable function. Please write down the first three terms of its Taylor expansion at point $x = 1$.

   $$P(x) = f(a) + f'(a)(x-1) + \frac{f''(x)}{2}(x-1)^2$$

   Notes: A Taylor series is a series that is used to create an estimate (guess) of what a function looks like. There is also a special kind of Taylor series called a Maclaurin series. The theory behind the Taylor series is that if a point is chosen on the coordinate plane (x- and y-axes), then it is possible to guess what a function will look like in the area around that point. This is done by taking the derivatives of the function and adding them all together. The idea is that it is possible to add the infinite number of derivatives and come up with a single finite sum. A Taylor series shows a function as the sum of an infinite series. The sum's terms are taken from the function's derivatives. A Taylor series for a function f at a point a looks like:

   $$f(x-a) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

   Here, we are getting the exact solution at that point because we are performing an infinite sum. This cannot be done on the computer, where instead we approxiamte that by trucating the summation at a particular n.

3. For the infinite sum $\sum_{n=1}^{\infty} \frac{1}{n^\alpha}$, where α is a positive real number, give the exact range of α such that the series converges.

   $$\alpha > 1$$

Notes: this is a p-series which is a special case of a harmonic series. The harmonic series looks like this:

$$\sum_{n=0}^{\infty} \frac{1}{n}$$

The harmonic (harmonic because of overtones) series is just a summation of infinite positive terms so it diverges. The p-series is a more general form of the harmonic series. If p is greater than 1 it will converge because we are adding more or less adding zero's after a certain point which is the point to which the series will converge to.

2 Linear algebra

1. What is the eigen-decomposition of a real symmetric matrix $A_{n\times n}$? Write down one form of that decomposition and explain each term in your formula. Based on these terms, suppose all eigenvalues are positive, derive $A^{-1/2}$.

   Eigen-decomposition can only be done on square matricies that are diaganolizable. Square matricies have the same rows and colums. If $A_{n\times n}$ is not symmetric, but has n linearly independent eigen vectors, we say the matrix is diagnolizable. A diaganolizable matrix is a matrix that can be represented as diagnol matrix. A diagnol matrix has non-zero entries only along its diagnol:

   $$S^{-1}AS = D$$

   where $S \in R^{n*n}$ that has the eigenvectors of A and $D \in Rn * n$ and is a diaganol matrix that contains the eigenvalues of A along its diaganol. Eigen-decomposition just rearranges this equation to yield the following:

   $$A = SDS^{-1}$$

   In the case of A being a symmetric, we know that it will have n real and n orthogonal eigenvectors, meaning it is diaganolizable. Furthermore, since it has orthognal eigen vectors, $S^{-1} = S^T$. Therefore, the eigen-decomposition becomes:

   $$A = SDS^T$$

   This technique can be handy to calculate the power of a matrix. In general,

   $$A^n = SD^nS^{-1}$$

   Therefore,

   $$A^{-1/2} = SD^{-1/2}S^T$$

2. What is a symmetric positive definite matrix $A_{n\times n}$? Give one of the equivalent definitions and explain your notation.

   Because the matrix is symmetric, we can say that $A = A^T$. Symmetrix matricies are seen a lot in machine learning (covariance, hessians). In order to understand what it means to be positive definite, we look at the quadratic form of the matrix. Explicitly, given the matrix A and a vector $x \in R^n$, the scalar quantatity as a result of the following operation:

   $$xA^Tx$$

is the quadratic form of the matrix. This quantatiy is a calculating a dot product between the input x and the output Ax. Now if a matrix has a quadratic form $> 0, \forall x$, the matrix can be called positive definite. If this value is $\geq 0, \forall x$, the matrix is positive semi-definite. If this value is $\leq 0, \forall x$, the matrix is negative semi-definite. If a matrix has a quadratic form $< 0, \forall x$, the matrix can be called negative definite. The defiteness of a matrix helps us in determining the convexity of an operation or of a loss function. A positive definite matrix is convex, meaning we can optimize it with iterative algorithms (gradient descent).

3. True/False. If you claim a statement is false, explain why. For two real matrices $A_{m\times n}$ and $B_{n\times m}$

   (a) Rank$(A) = \min\{m, n\}$

   False, Rank$(A) \leq \min\{m, n\}$

   (b) If $m = n$, then trace$(A) = \sum_{i=1}^{n} A_{ii}$

   True

   (c) If A is a symmetric matrix, then all eigenvalues of A are real

   True

   (d) If A is a symmetric matrix, λ_1 and λ_2 are two distinct eigen-values and v_1,v_2 are the corresponding eigen-vectors, then it is possible that $v_1^T v_2 > 0$.

   False, the eigen-vectors of a symmetric matrix are orthogonal, meaning $v_1^T v_2 = 0$

   (e) If A is a symmetric matrix, v_1,v_2 are two distinct eigen-vectors of λ, then it is possible that $v_1^T v_2 > 0$.

   False, the even if eigen-values are not distinct, eigen-vectors of a symmetric matrix are orthogonal, meaning $v_1^T v_2 = 0$

   (f) trace(ABAB) = trace(AABB)

   False, you cannot even compute the RHS because AA has mismatched dimensions ($n \neq m$).

3 Statistics

1. X_1, X_2, ..., X_n are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ random variables, where $\mu \in \mathbb{R}$ and $\sigma > 0$ is finite. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

   (a) What is an unbiased estimator? Is $\bar{X}_n$ an unbiased estimator of μ?

   The bias of an estimator is how different the expected value of the estimator is from the true value of the parameter being estimated. More concretely, say you are estimating a parameter, θ with an esitmator $\hat{\theta}$. Then the bias of the estimator is calculated as follow:

   $$Bias(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta$$

   An estimator is said to be unbiased if the bias equals 0 or,

   $$E[\hat{\theta}] = \theta$$

Intuitively, we might want this to be a property of our estimator because it means in the long run (expectation) it is doing a good job estimating our paramater. As far as the case where we are using $\bar{X}_n$ (sample mean) as an estimator of μ (population mean), the sample mean is an unbiased estimator of the population mean because

$$E(\bar{x}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

(b) What is $E[(\bar{X}_n)^2]$ in terms of n, μ, σ?

$$var(\bar{x}) = E[(\bar{X}_n)^2] - E[\bar{X}_n]^2 E[(\bar{X}_n)^2] = var(\bar{x}) + E[\bar{X}_n]^2 E[(\bar{X}_n)^2] = \frac{\sigma^2}{n} + \mu^2$$

(c) Give an unbiased estimator of σ^2.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

This is almost the intuitive definition of the sample variance. The only thing a little weird is the $\frac{1}{n-1}$ normalization constant. This can be understood as a correction factor to ensure the estimator is a good/intuitive estimator, as well as unbiased.

(d) What is a consistent estimator? Is $\bar{X}_n$ a consistent estimator of μ?

An estimator is called consistent when as the number of data points (n) used to generate the estimator increases indefinitely, the estimator approaches the true value of the paramter. More concretely, say you are estimating a parameter, θ with an esitmator $\hat{\theta}$. Then the estimator is a consistent one if $Var(\hat{\mu}) \to 0$ as $n \to \infty$. The sample mean calculated this way is consistent because it is itself normally distributed with a mean μ and a varaince $\frac{\sigma^2}{n}$. The variance here clearly approaches inifinity as n does, therefore it is consistent.

2. Suppose $X_{p\times 1}$ is a vector of covariates, $\beta_{p\times 1}$ is a vector of unknown parameters, ϵ is the unobserved random noise and we assume the linear model relationship $y = X^T\beta + \epsilon$. Suppose we have n i.i.d. samples from this linear model, and the observed data can be written using the matrix form: $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}$.

(a) If we want to estimate the unknown β using a least square method, what is the objective/loss function $L(\beta)$ to obtain $\widehat{\beta}$?

The loss for the i^{th} example is

$$L(\beta) = (\mathbf{X}^{(i)} - \mathbf{y}^{(i)})^2$$

where $\mathbf{X}^{(i)}$ is the i^{th} row of $\mathbf{X}$ and $\mathbf{y}^{(i)}$ is the i^{th} row of $\mathbf{y}$. This is the L-2 loss common in regression problems.

(b) What is the solution of $\widehat{\beta}$? Represent the solution using the observed data $\mathbf{y}$ and $\mathbf{X}_{n\times p}$. Note that you may assume that $\mathbf{X}^T\mathbf{X}$ is invertible.

To solve, we take our cost function, $J(\beta) = (\mathbf{X}\beta - \mathbf{y})^T(\mathbf{X}\beta - \mathbf{y})$ (the cost is just the loss averaged over all examples), and find its gradient with respect to β. Remember that our cost is a function that takes in a matrix and returns a scalar ($J(\theta) : R^n \to R$):

$$\nabla_\theta J(\beta) = \nabla_\beta \frac{1}{2}(\mathbf{X}\beta - \mathbf{y})^T(X\beta - \mathbf{y}) \tag{1}$$
$$= \frac{1}{2}\nabla_\theta((\mathbf{X}\beta)^T\mathbf{X}\beta - (\mathbf{X}\beta)^T\mathbf{y} - \mathbf{y}^T(\mathbf{X}\beta) + \mathbf{y}^T\mathbf{y}) \tag{2}$$
$$= \frac{1}{2}\nabla_\beta(\beta^T(\mathbf{X}^T\mathbf{X})\beta - \mathbf{y}^T(\mathbf{X}\beta) - \mathbf{y}^T(\mathbf{X}\beta)) \tag{3}$$
$$= \frac{1}{2}\nabla_\beta(\beta^T(\mathbf{X}^T\mathbf{X})\beta - 2(\mathbf{X}^T\mathbf{y})^T\beta) \tag{4}$$
$$= \frac{1}{2}(2\mathbf{X}^T\mathbf{X}\beta - 2\mathbf{X}^T\mathbf{y}) \tag{5}$$
$$= \mathbf{X}^T\mathbf{X}\beta - \mathbf{X}^T\mathbf{y} \tag{6}$$

Setting this equal to 0 gives us the normal equations:

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}$$

Our optimal beta ($\widehat{\beta}$):

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

4 Programming

1. Use the following code to generate a set of n observations $\mathbf{y}_{n\times 1}$ and $\mathbf{X}_{n\times p}$. Follow the previously established formula to solve for the least square estimator $\widehat{\beta}$. Note that you must write your own code, instead of using existing functions such as `lm()`. In addition, what should you do if you are asked to add an intercept term β_0 into your estimation (even the true $\beta_0 = 0$ in our data generator)?

```
set.seed(1)
n = 100; p = 5
X = matrix(rnorm(n * p), n, p)
y = X %*% c(1, 0, 0, 1, -1) + rnorm(n)
```

[1]:
```python
import numpy as np
from numpy.linalg import inv

# reproduced the R code provided to generate the data in Python
def generate(n, p, beta):
    np.random.seed(0)
    X = np.random.normal(size=(n,p))
    np.random.seed(0)
    noise = np.random.normal(size=n)
    y = X @ beta + noise
    return y, X, beta

def solve(X, y):
```

```
    return inv(X.T@X)@X.T@y

n = 100
p = 5
beta = np.array([1.0,0.0,0.0,1.0,-1.0])

(y,X,beta) = generate(n, p, beta)
beta_hat = solve(X, y)
print(f'beta (true): {beta}')
print(f'beta_hat (estimate): {beta_hat}')
print(f'Average squared difference between beta (true) and beta_hat (estimate):␣
 ↪{np.mean(np.power(beta-beta_hat,2))}')
```

```
beta (true): [ 1.  0.  0.  1. -1.]
beta_hat (estimate): [ 1.01740332  0.09935171 -0.03501228  0.88735656
-0.74824799]
Average squared difference between beta (true) and beta_hat (estimate):
0.01749342291415676
```

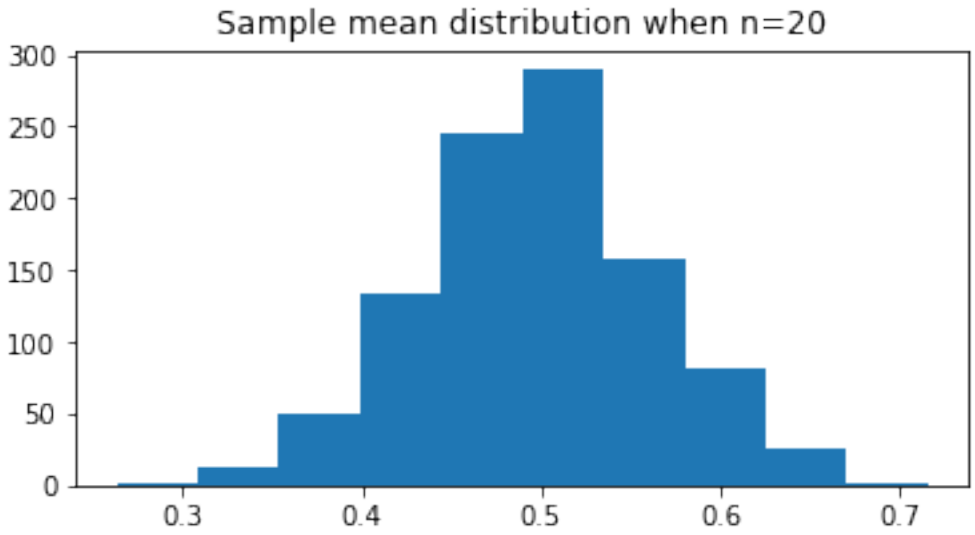
In order to add an intercept term, we append a 1 to each row of $\mathbf{X}$. This makes $X \in R^{n\times(p+1)}$. Now the same code gives us a beta vector containing 6 paramters, with the last one being our estimate of β_0.

2. Perform a simulation study to check the consistency of a sample mean estimator $\bar{X}_n$. Please save your random seed so that the results can be replicated by others.
   (a) Generate a set of $n = 20$ i.i.d. observations from uniform(0, 1) distribution and calculate the sample mean $\bar{X}_n$
   (b) Repeat step (a) 1000 times to collect 1000 such sample means and plot them using a histogram.
   (c) How many of such sample means (out of 1000) are at least 0.1 away from true mean parameter, which is 0.5 for uniform (0, 1)?
   (d) Repeat steps (a) to (c) with $n = 100$ and $n = 500$. What conclusion can you make?

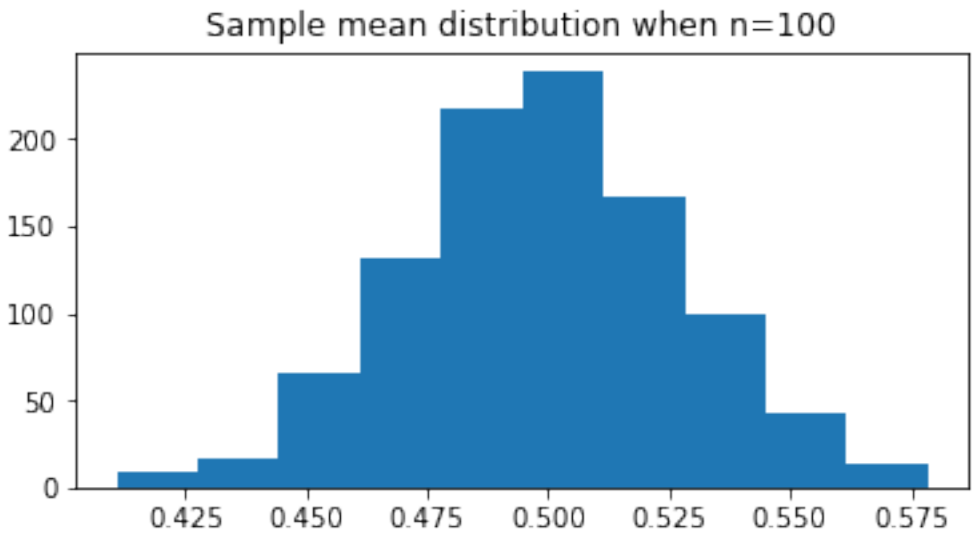
[28]:
```
import numpy as np
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [6, 3]

def step_a(n):
    return np.mean(np.random.uniform(size=n))

def step_b(n):
    sample_means = []
    for i in range(1000):
        sample_means.append(step_a(n))
    plt.hist(sample_means)
    plt.title(f'Sample mean distribution when n={n}')
    plt.show()
    return sample_means
```

```
def step_c(sample_means):
    return (np.abs(np.array(sample_means)-0.5) > 0.1).sum()

# step d
ns=[20,100,500]
for n in ns:
    print(f'When n = {n}')
    sample_means = step_b(n)
    print(
        f'There are {step_c(sample_means)} at least 0.1 away from true mean␣
 ↪parameter'
        f'when n={n}, which is 0.5 for uniform (0, 1)\n'
    )
```

```
When n = 20
```

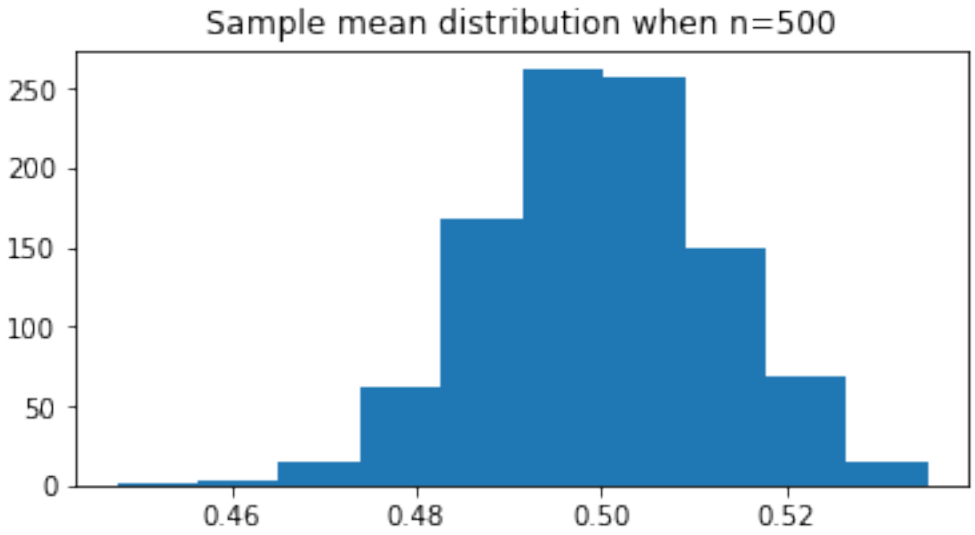



```
There are 134 at least 0.1 away from true mean parameterwhen n=20, which is 0.5
for uniform (0, 1)

When n = 100
```




```
There are 0 at least 0.1 away from true mean parameterwhen n=100, which is 0.5
for uniform (0, 1)
```

```
When n = 500
```




```
There are 0 at least 0.1 away from true mean parameterwhen n=500, which is 0.5
for uniform (0, 1)
```

As we increase the number of samples, the variance of the distribution of the sample means gets tighter. We can see this in this histograms by looking at how spread out the x-axis is. For n=20,

the histogram is most spread and when n=500, the histogram is tightest. We can also see this by calculating how many samples in the distribution are at least 0.1 away from the true mean paramter, which is 0.5. This is largest when n=20 and 0 when n=100 or n=500. This makes sense because as stated above, the sample mean is a consistent estimator, ie its variance becomes 0 as you increase the sample size.