

Homework 1

● Graded

Student

John Stephen Gutam

Total Points

71 / 100 pts

Question 1

Fair coins

Resolved 18 / 25 pts

✓ + 25 pts Correct

+ 7 pts Correct histograms for v_1 and v_r

+ 3 pts Correct histograms for v_{min}

+ 10 pts Correct figure for the Hoeffding bound

+ 5 pts Correct explanations for the trend in Hoeffding bound graph.

– 7 pts Wrong explanation of the Hoeffding bound. Also missing are curves for v_1 and v_r in the bound graph [-7 points]

C Regrade Request

Submitted on: Feb 26

Dear Gaurab,

I request you to please regrade my question 1 for the graphs and coding. It looks similar to the solutions that have been provided.

Thank you.

The one graph you have provided in the writeup are for that of v_{min} -- which is correct. Again, as the comment already points out you are missing graphs for v_1 and v_r [-5 points].

An additional two points were deducted because your rationale as to why the bound is not followed by v_{min} is incorrect or rather missing. Check the answer key for the solution, but its essentially because you are picking the minimum after looking at the data.

Reviewed on: Feb 27

Question 2

Perceptron learning

17 / 30 pts

+ 30 pts Correct

✓ + 15 pts Correct implementation

✓ + 10 pts Histograms look correct

✓ + 5 pts Correct characterization of longer iterations with appropriate evidence to back up claim

💬 - 13 pts Missing analysis of when the algorithm takes more iterations [-5 points]

Graph is off. Should be centered around approximately ~12.5 (check answer key) for correct graph.

- 1 It would help in debugging if we could look at the graph of just the iterations. The problem writeup also explicitly asks for this graph. [-8 points]

Question 3

Chebyshev's inequality

15 / 15 pts

✓ + 15 pts Correct

+ 5 pts Correct Part (a)

+ 5 pts Correct Part (b)

+ 5 pts Correct Part (c)

Question 4

Hyperplanes

Resolved 6 / 10 pts

+ 10 pts Correct

✓ + 5 pts Correct values of $a = \frac{-w_1}{w_2}$ and $b = \frac{-w_0}{w_2}$ for part (a)

+ 5 pts Correct part (b) with a properly shown figure.

💬 + 1 pt Your lines are wrong. The dividing line should be the same in both cases with only the positive and negative regions being flipped between the two cases [+1 point] for the attempt.

C Regrade Request

Submitted on: Feb 27

For question 4: I request you to regrade the question 4. My graph exactly matches with the solution document which is a decreasing one and also why do the 2 vectors need to be the same?

One of your lines match the solution key. the other one does not. Both lines are supposed to be the same because when you plug in the numbers, the slope and the intercept are the same i.e. your line for $w = (1, 2, 3)$ should look like your graph for $w = (-1, -2, -3)$. The difference between the two hyperplanes is which regions they designate as being positive and negative.

Reviewed on: Feb 27

Question 5

Gaussian log likelihood

10 / 10 pts

✓ + 10 pts Correct

+ 2 pts Recognize the need to take partial derivatives with respect to μ and σ^2

+ 4 pts Correct partial derivative with respect to μ

+ 4 pts Correct partial derivative with respect to σ^2

Question 6

Likelihood and sigmoids

Resolved

5 / 10 pts

+ 10 pts Correct

+ 7 pts Most of the way there, some details missing or incorrect.

✓ + 5 pts About half of the problem is solved. Some correct argumentation presented.

+ 3 pts The solution started in the right direction. However no significant progress made.

+ 1 pt Attempt at a solution

+ 0 pts Missing/Incorrect solution

Derivation of the first part ~~missing~~ incomplete. In the second part, if you add y_i to the exponent, you can get rid of the indicator function to get a unified expression. Check the answer key for the correct solution.

C Regrade Request

Submitted on: Feb 26

Dear Gaurab,

I request you to please regrade my question 6 - part (a). It had been given 0 marks for it.

Question 6:

I defined the likelihood function with $L(\Theta)$ in terms $h(x_i)$ and probability $P(y=+1 | x)$ from where it has been defined for the +1 y value and -1 y value.

Thank you.

Oh, that comment is an artifact of me re-using comments from other assignments. In the first part, you the quantities inside the log are supposed to be fractions. You are short a couple of steps from reaching the statement given in the homework statement. Its not missing, its incomplete.

I have awarded points overall for carrying through with the derivation of the second part (which is also incomplete and is actually the expression that needs to be **maximized** not minimized -- as the question asks). I apologize for this confusion, but you have already been given the maximum points possible on this.

Reviewed on: Feb 27

Question assigned to the following page: [1](#)

Name: John Stephen Gutam
Email: jgutam@gmu.edu

Problem 1:

Plotting the histogram of the distribution nu1, nu_r, nu_min over the 100000 samples.

In this experiment of flipping 1000 different fair coins 10 times each and calculating the proportions of heads for the first coin, a randomly chosen coin, and the coin that came up heads the least among all of them.

The coin_experiment() code uses the numpy library to generate random choices for each coin flip, and then calculates the proportion of heads for each coin. It returns the proportions for the first coin (nu1), a randomly chosen coin (nu_r), and the coin that came up heads the least (nu_min).

```
In [7]: # import numpy as np
def coin_experiment():
    num_coins = 1000
    num_flips = 10

    # Initialize arrays to store the results
    results = np.zeros((num_coins, num_flips))
    proportions = np.zeros((num_coins))

    for i in range(num_coins):
        # Flip the coin num_flips times
        flips = np.random.choice([0, 1], size=num_flips)
        results[i] = flips

        # calculate the proportion of heads
        proportion_heads = np.mean(flips)
        proportions[i] = proportion_heads

    # calculate the proportion of heads for the first coin
    nu1 = proportions[0]

    # calculate the proportion of heads for a randomly chosen coin
    nu_r = np.random.choice(proportions)

    # calculate the proportion of heads for the coin that came up heads the least
    nu_min = np.min(proportions)

    return nu1, nu_r, nu_min
```

Next, we repeated the entire experiment 100,000 times and plotted the histograms of the distribution of nu1, nu_r, and nu_min. Below is the output.

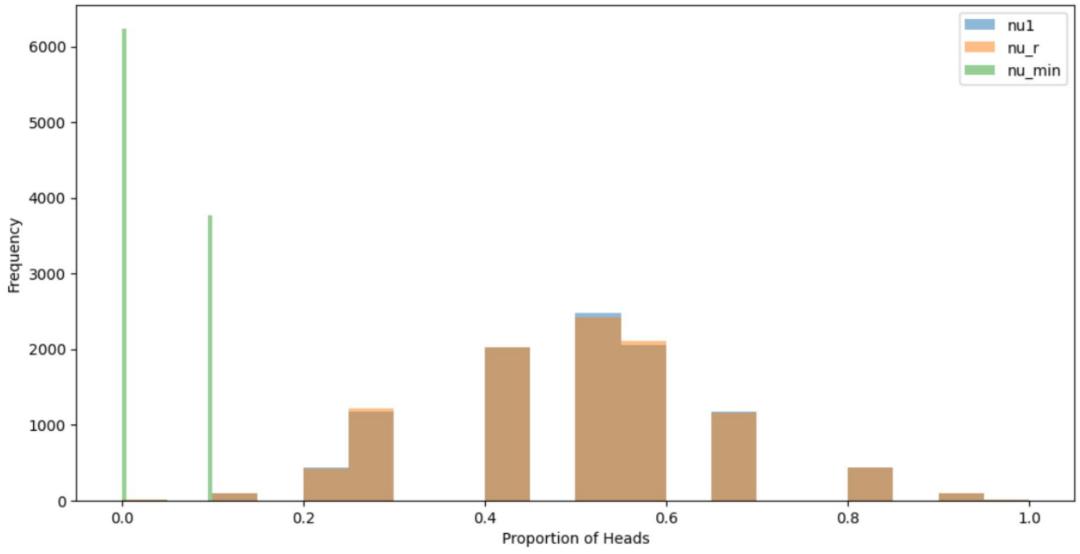
```
In [9]: # import matplotlib.pyplot as plt
def repeat_experiment(num_repeats):
    nu1_values = np.zeros((num_repeats))
    nu_r_values = np.zeros((num_repeats))
    nu_min_values = np.zeros((num_repeats))

    for i in range(num_repeats):
        nu1, nu_r, nu_min = coin_experiment()
        nu1_values[i] = nu1
        nu_r_values[i] = nu_r
        nu_min_values[i] = nu_min

    # Plot the histograms
    plt.figure(figsize=(12, 6))
    plt.hist(nu1_values, bins=20, alpha=0.5, label='nu1')
    plt.hist(nu_r_values, bins=20, alpha=0.5, label='nu_r')
    plt.hist(nu_min_values, bins=20, alpha=0.5, label='nu_min')
    plt.xlabel('Proportion of Heads')
    plt.ylabel('Frequency')
    plt.legend()
    plt.show()

repeat_experiment(10000)
```

Question assigned to the following page: [1](#)



Observation from the histogram1:

The range of possible proportions of heads is divided into intervals, called bins. As we are flipping a coin 10 times, the proportion of heads ranges from 0 to 1 in increments of 0.1. Each interval represents a possible outcome of the experiment.

For each interval, we counted how many times the proportion of heads occurred in our experiment. This count represents the frequency of occurrence for that proportion.

The histogram plot consists of bars, where the height of each bar represents the frequency of occurrence of the corresponding proportion of heads. The x-axis represents the range of possible proportions, while the y-axis represents the frequency of occurrence.

By looking at the histogram we can assess the distribution of the observed proportion of heads(observed empirical risk) for random nu_r is more concentrated at 0.5 and even for the first sample it is showing at 0.5. The proportions are more common at 0.5. Hence the distribution of proportions of heads aligns with our expectations compared to the true population proportion ($\nu = 0.5$)

Question assigned to the following page: [1](#)

Plotting the histogram for estimates of $\Pr(|v-u| > \epsilon)$ as a function of epsilon.

We calculated the estimates of $\Pr(|v-u| > \epsilon)$ as a function of epsilon and plotted them together with the Hoeffding bound. It uses the coin_experiment function to generate the nu_min values and checks if they exceed epsilon to update the estimates

```
def calculate_hoeffding_bound(epsilon, num_flips):
    return 2 * np.exp(-2 * epsilon**2 * num_flips)

def plot_hoeffding_bound(num_repeats):
    epsilons = np.linspace(0, 0.5, 100)
    hoeffding_bounds = calculate_hoeffding_bound(epsilons, num_flips=10)

    nu_values = np.zeros((num_repeats))
    hoeffding_estimates = np.zeros_like(epsilons)

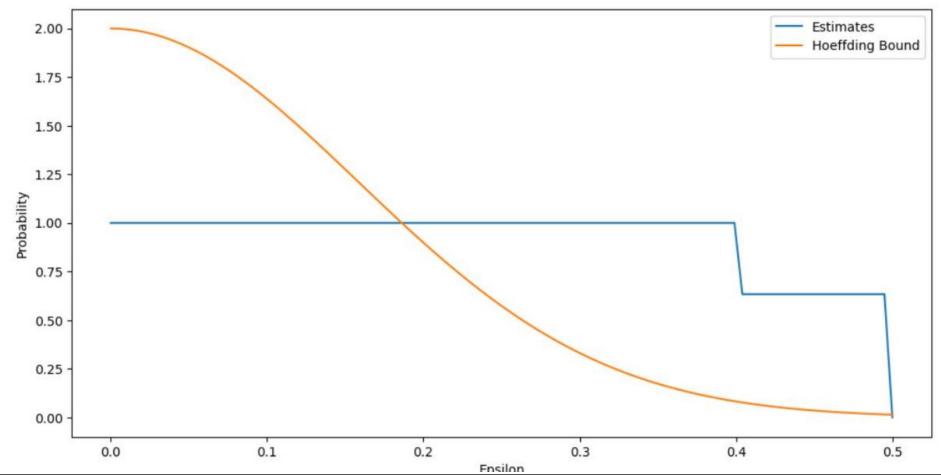
    for i in range(num_repeats):
        nu1, nu_r, nu_min = coin_experiment()
        nu_values[i] = nu_min

        for j, epsilon in enumerate(epsilons):
            if abs(nu_min - 0.5) > epsilon:
                hoeffding_estimates[j] += 1

    # Normalize the estimates
    hoeffding_estimates /= num_repeats

    # Plot the estimates and the Hoeffding bound
    plt.figure(figsize=(12, 6))
    plt.plot(epsilons, hoeffding_estimates, label='Estimates')
    plt.plot(epsilons, hoeffding_bounds, label='Hoeffding Bound')
    plt.xlabel('Epsilon')
    plt.ylabel('Probability')
    plt.legend()
    plt.show()

plot_hoeffding_bound(10000)
```



Observations from the histogram2:

The observed proportion of heads (v) and the true proportion of heads ($u = 0.5$), the histogram is plotted for $\Pr(|v-u| > e)$ as a function of e .

The range of possible deviations (e) is divided into intervals that range from 0 to 0.5 into smaller intervals, each representing a different magnitude of deviation.

The observed proportion of heads for a coin (v) falling within the Hoeffding bound (as most concentration in the histogram is on 0.5), it means that deviation from the true population (u) is within the acceptable range. In other words, these coins obey the Hoeffding bound, and the observed proportions of heads are consistent with what would be expected ($u = 0.5$) based on random chance.

Also, for more samples, the whole epsilon term decreases therefore the probability also decreases.

Question assigned to the following page: [3](#)

Problem 3:

G. John Stephen
jgutam@gmu.edu

Problem 3

a) If t is a non-negative random variable, for any $\alpha > 0$, $P_r(t \geq \alpha) \leq E[t]/\alpha$.

Proof:

$$P_r(t \geq \alpha) = \int_{\alpha}^{\infty} f(t) dt \quad (\text{where } f(t) \text{ is the probability density function of } t) \rightarrow \text{eq ①}$$

Now, condition the expectation on the event $t \geq \alpha$:

$$E[t] = \int_0^{\infty} t \cdot f(t) dt = \int_{\alpha}^{\infty} t \cdot f(t) dt + \int_0^{\alpha} t \cdot f(t) dt.$$

since t is non negative:

$$E[t] \geq \int_{\alpha}^{\infty} t \cdot f(t) dt \geq \alpha \int_0^{\infty} f(t) dt = \alpha E[t] \quad (\text{From eq ①})$$

$$E[t] \geq \alpha P_r(t \geq \alpha)$$

Divide by α :

$$P_r(t \geq \alpha) \leq \frac{E[t]}{\alpha} \rightarrow \text{eq ②}$$

b) If u is any random variable with mean μ and variance σ^2 , and $\alpha > 0$, then $P_r((u-\mu)^2 \geq \alpha) \leq \frac{\sigma^2}{\alpha}$

Proof:

Apply the non-negative random variables $(u-\mu)^2$ with α

$$P_r((u-\mu)^2 \geq \alpha) \leq \frac{E[(u-\mu)^2]}{\alpha} \quad (\because \text{eq ②})$$

Now, note that $E[(u-\mu)^2] = \text{Var}(u) = \sigma^2$

$$\therefore P_r((u-\mu)^2 \geq \alpha) \leq \frac{\sigma^2}{\alpha}$$

Question assigned to the following page: [3](#)

Part (c) :

If u_1, \dots, u_n are independent and identically distributed, each with mean μ and variance σ^2 , and \bar{u} is the sample mean of u_1, \dots, u_n , then for

$$\alpha > 0 : P_{\lambda}((\bar{u} - \mu)^2 \geq \alpha) \leq \frac{\sigma^2}{n\alpha}$$

Proof:

Apply part (b) to the random variable \bar{u} with α and note that

$$\text{Var}(\bar{u}) = \frac{\sigma^2}{n} \text{ for } n \text{ iid means } \mu \text{ and variance } \sigma^2$$

$$P_{\lambda}((\bar{u} - \mu)^2 \geq \alpha) \leq \frac{\text{Var}(\bar{u})}{\alpha} = \frac{\sigma^2}{n\alpha}$$

$$(0 \leq \alpha \leq \infty) \Rightarrow (\lambda \leq \infty) \Rightarrow (\lambda \leq \infty)$$

$$(\alpha \leq 0) \Rightarrow (\lambda \leq 0)$$

$$\text{Op} \rightarrow \boxed{\frac{(\lambda)^2}{\alpha} \geq (\lambda \leq 0)^2}$$

$$\text{Op} \rightarrow (\lambda \leq 0)^2 \geq (\lambda \leq 0)^2$$

$$\text{Op} \rightarrow (\lambda \leq 0)^2 \geq (\lambda \leq 0)^2$$

$$\text{Op} \rightarrow (\lambda \leq 0)^2 \geq (\lambda \leq 0)^2$$

$$\text{Op} \rightarrow (\lambda \leq 0)^2 \geq (\lambda \leq 0)^2$$

$$\boxed{\lambda \geq 0}$$

Question assigned to the following page: [4](#)

Problem 4:

4) Problem 4:

Part 1

The perceptron prediction is given by $h(x) = \text{sign}(\omega \cdot x)$ where $\omega = (\omega_0, \omega_1, \omega_2)$ and $x = \underbrace{(1, x_1, x_2)}_{\text{homogeneous}}$. The regions on the plane where $h(x) = +1$ and $h(x) = -1$ are separated by line.

If we set $h(x) = 0$, we can find the equation of the decision boundary.

$$h(x) = \text{sign}(\omega \cdot x)$$

$$0 = \omega \cdot x$$

$$0 = \omega_0 \cdot 1 + \omega_1 \cdot x_1 + \omega_2 \cdot x_2$$

Solving for

$$x_2 = -\frac{\omega_0}{\omega_2} - \frac{\omega_1}{\omega_2} x_1, \quad \begin{array}{l} \text{(If it is in the form of a} \\ \text{line, } ax + b = y \end{array}$$

(we compare the
coefficients.)

$$a = -\frac{\omega_1}{\omega_2}, \quad b = \frac{-\omega_0}{\omega_2}$$

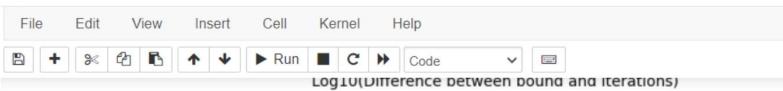
So, the values of a and b in terms of the elements of ω are $a = -\frac{\omega_1}{\omega_2}$

$$\text{and } b = \frac{-\omega_0}{\omega_2}$$

Part (b)

The plot will show two lines, one for each weight vector. The lines separate the regions where $h(x) = +1$ and $h(x) = -1$. The lines have opposite slopes and intercepts, illustrating the difference between the two weight vectors of their signs and magnitude.

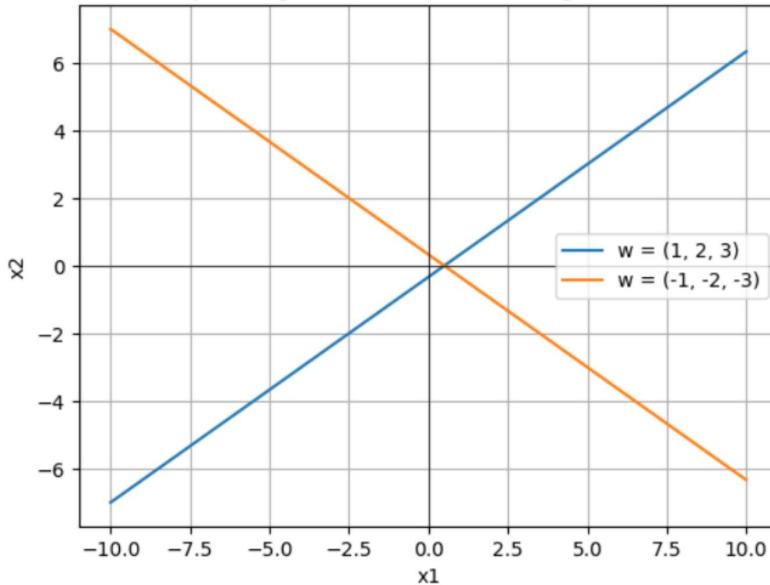
Question assigned to the following page: [4](#)



1 # Problem 4

```
In [2]: M 1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 # Weight vector w = (1, 2, 3)
5 w1 = 2/3
6 w2 = -1/3
7 x1_vals = np.linspace(-10, 10, 100)
8 x2_vals = w1*x1_vals + w2
9
10 plt.plot(x1_vals, x2_vals, label='w = (1, 2, 3)')
11
12 # Weight vector w = (-1, -2, -3)
13 w1 = -2/3
14 w2 = 1/3
15 x2_vals = w1*x1_vals + w2
16
17 plt.plot(x1_vals, x2_vals, label='w = (-1, -2, -3)')
18
19 plt.xlabel('x1')
20 plt.ylabel('x2')
21 plt.axhline(0, color='black', linewidth=0.5)
22 plt.axvline(0, color='black', linewidth=0.5)
23 plt.legend()
24 plt.grid(True)
25 plt.title('Separating Lines for Different Weight Vectors')
26 plt.show()
```

Separating Lines for Different Weight Vectors



Question assigned to the following page: [5](#)

Problem 5:

G. John Stephen.

⑤ Problem 5

To derive the estimates μ and σ^2 that maximize the log-likelihood function for a given distribution, we will find the maximum by taking partial derivatives with respect to μ and σ^2 , setting them to zero, and solving for the parameters.

Estimating μ :

$$\ln P(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi)$$

partial derivative w.r.t μ

$$\frac{\partial}{\partial \mu} \ln P(x|\mu, \sigma^2) = -\frac{1}{\sigma^2} \times 2 \sum_{i=1}^n (x_i - \mu) - 0 - 0 = 0$$

$$= -\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^n (x_i - \mu) = 0$$

$$\sum_{i=1}^n x_i - n\mu = 0 \Rightarrow \mu = \frac{1}{n} \sum_{i=1}^n x_i$$

so the maximum likelihood estimate for μ is the sample mean

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

Estimating σ^2 :

$$\ln P(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi)$$

partial derivative w.r.t σ^2

$$\frac{\partial}{\partial \sigma^2} \ln P(x|\mu, \sigma^2) = -\frac{1}{2} (-2)\sigma^{-3} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \frac{1}{\sigma^2} (2\sigma) - 0$$

Question assigned to the following page: [5](#)

$$-\frac{1}{2} (\sigma^2) - \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \frac{1}{\sigma^2} (\sigma^2) = 0$$

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = \frac{n}{\sigma^2}$$

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = n$$

$$\therefore \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

Question assigned to the following page: [6](#)

Problem 6:

6. Problem 6

The likelihood function L measures the likelihood of observing the given data $\{(x_i, y_i)\}_{i=1}^n$, given the parameters θ of the model.

We model the probability as $P(y = +1 | x)$ using a function $h(x)$.

The likelihood function for a single data point (x_i, y_i) is given by:

$$L(\theta) = \begin{cases} h(x_i) & \text{if } y_i = +1 \\ 1 - h(x_i) & \text{if } y_i = -1 \end{cases}$$

$$l(\theta) = \sum_{i=1}^n [I[y_i = +1] \log(h(x_i)) + I[y_i = -1] \log(1 - h(x_i))]$$

where $I[A]$ is the indicator function.

Given $h(x) = \sigma(\omega \cdot x)$, where $\sigma(z)$ is the sigmoid function.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Substitute

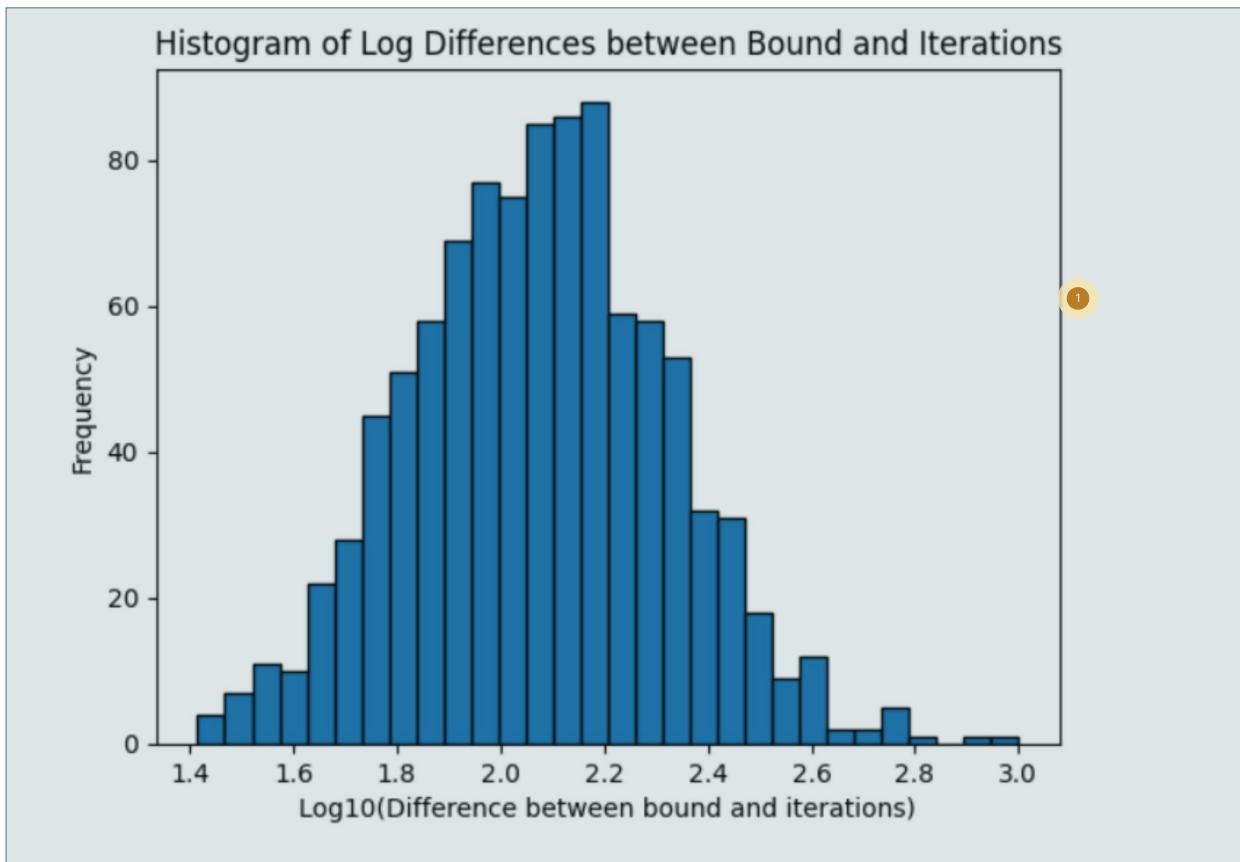
$$\begin{aligned} l(\theta) &= \sum_{i=1}^n [I[y_i = +1] \log(\sigma(\omega \cdot x_i)) + I[y_i = -1] \log(1 - \sigma(\omega \cdot x_i))] \\ &= \sum_{i=1}^n [I[y_i = +1] \log\left(\frac{1}{1 + e^{-\omega \cdot x_i}}\right) + I[y_i = -1] \log\left(1 - \frac{1}{1 + e^{-\omega \cdot x_i}}\right)] \\ &= \sum_{i=1}^n [I[y_i = +1] \log\left(\frac{1}{1 + e^{-\omega \cdot x_i}}\right) + I[y_i = -1] \log\left(\frac{e^{-\omega \cdot x_i}}{1 + e^{-\omega \cdot x_i}}\right)] \\ &= \sum_{i=1}^n [I[y_i = +1] \log\left(\frac{1}{1 + e^{-\omega \cdot x_i}}\right) + I[y_i = -1] \log\left(\frac{1}{1 + e^{\omega \cdot x_i}}\right)] \\ &= \sum_{i=1}^n [I[y_i = +1] \log\left(\frac{1}{1 + e^{-\omega \cdot x_i}}\right) - I[y_i = -1] \log\left(1 + e^{\omega \cdot x_i}\right)] \end{aligned}$$

This expression represents the minimized when $h(x) = \sigma(\omega \cdot x)$.

Question assigned to the following page: [2](#)

Problem 2:

1. Generated an 11-dimensional weight vector w^* , where the first dimension is 0 and the other 10 dimensions are sampled independently at random from the uniform (0, 1) distribution.
2. Repeated the steps 1000 times:
 - a. Generated a random training set with 100 examples, where each dimension of each training example is sampled independently at random from the uniform (-1, 1) distribution, and the examples are all classified by w^* .
 - b. By running the perceptron learning algorithm on the training set, starting with the zero-weight vector, and keeping track of the number of iterations it takes to learn a hypothesis that correctly separates the training data.
3. Plotted a histogram of the number of iterations the algorithm takes to learn a linear separator.
4. Calculate the difference between the number of iterations and the bound on the number of errors derived in class for each experiment.
5. Plot a histogram of the logarithm of this difference.



The code generated a new weight vector and training set for each experiment. By running the perceptron learning algorithm on the training set and keeping track of the number of iterations required to learn a hypothesis that correctly separates the training data. It then plots a histogram of the number of iterations and calculates the difference between the number of iterations and the bound on the number of errors derived in class. Finally, plotted a histogram of the logarithm of this difference.

By analyzing the histograms, we examined the distribution of the number of iterations and the difference between the iterations and the bound. This analysis shows how the perceptron learning algorithm performs in different scenarios where the algorithm takes more iterations to correctly learn a hypothesis that separates the training data. The algorithm took more iterations between 2.0 and 2.2 to correctly separate the training data.

Question assigned to the following page: [2](#)

Note that the results may vary between different runs of the experiment due to the randomness involved in generating the weight vector and training set. Therefore, to run the experiment multiple times and analyzed the distribution of the results to draw meaningful conclusions.