ML Assignment 2

1 Illustration of Markov's, Chebyshev's and Hoeffding's Inequalities (23 points)

1.a

Derivation of bounds

For plotting the bounds, we derive explicit expressions of the bounds as a function of α and the bias, p, of $X_1, \cdots, X_{20}$. Let $M(\alpha, p), C(\alpha, p), H(\alpha, p)$ denote the Markov, Chebyshev and Hoeffding bound, respectively. Let $Y = \frac{1}{20}\sum_{i=1}^{n} X_i$. By linearity of expectation, we have that $\mathbb{E}(Y) = \frac{1}{20} \cdot 20p = p$. We have that

$$\mathbb{E}(X_1^2) = 1 \cdot p + (1-p) \cdot 0 = p$$

And thus

$$\mathrm{Var}(X_1) = E(X_1^2) - E(X_1)^2 = p - p^2 = p(1-p)$$

The variance of a sum of i.i.d. random variables is the sum of variances, and thus

$$\mathrm{Var(Y)} = \frac{1}{20^2}\sum_{i=1}^{n}(\mathrm{Var}(X_i)) = \frac{20p(1-p)}{20^2} = \frac{p(1-p)}{20}$$

For Markov's bound, we have that

$$M(\alpha, p) = \mathbb{P}(Y \geq \alpha) \leq \frac{E(Y)}{\alpha} = \frac{p}{\alpha}$$

For Chebyshev's bound, assume that $p < \alpha$, and that $C(\alpha, p) \leq 1$. If not, we set $C(\alpha, p) = 1$. We have that

$$C(\alpha, p) = \mathbb{P}(Y \geq \alpha) = \mathbb{P}(Y - \mathbb{E}(Y) \geq \alpha - \mathbb{E}(Y)) \leq \mathbb{P}(|Y - \mathbb{E}(Y)| \geq \alpha - \mathbb{E}(Y)) \leq \frac{\mathrm{Var}(Y)}{(\alpha - \mathbb{E}(Y))^2} = \frac{p(1-p)}{20(\alpha - p)^2}$$

For Hoeffdings inequality (in the form of corollary 2.5, which apply since X_i are Bernoulli R.V.'s and thus $X_i \in [0, 1]$ and $\mathbb{E}(X_i) = p$ for all i) we have that

$$H(\alpha, p) = \mathbb{P}(Y \geq \alpha) = \mathbb{P}(Y - \mathbb{E}(X_1) \geq \alpha - \mathbb{E}(X_1)) \leq e^{-2\cdot 20\cdot(\alpha - \mathbb{E}(X_i))^2} = e^{-40(\alpha - p)^2}$$

Plotting and granularity

As $\mathbb{P}(X_i \in \{0, 1\}) = 1$ for all i, we have that $\mathbb{P}(\sum_{i=1}^{20} X_i \in \{0, \cdots, 20\}) = 1$, and therefore that $\mathbb{P}(Y \in \{0, 0.05, 0.10, \cdots 0.9, 0.95, 1\}) = 1$. If we set $\alpha = 0.51$, we have that

$$\mathbb{P}(Y \geq \alpha) = \mathbb{P}(Y \geq 0.55)$$

Therefore adding, say, $\alpha = 0.51$ would not lead to different results to the expirements, and therefore would not provide any extra information. We plot the empirical frequency, and the probability bounds below

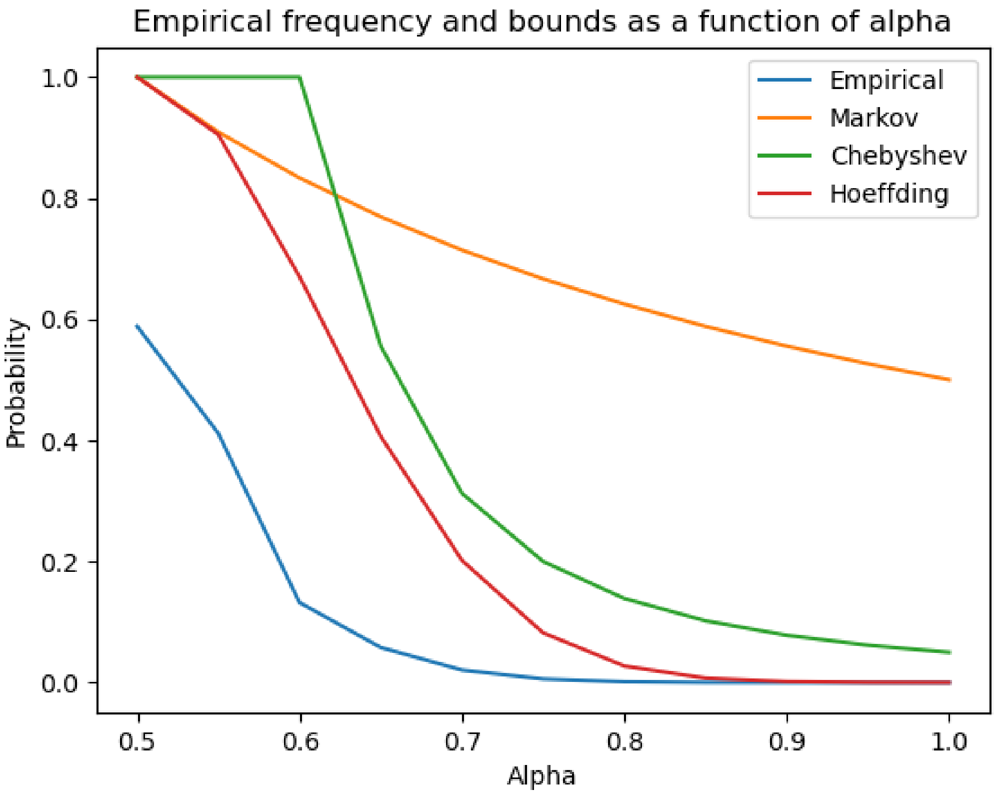

Figure 1: $p = \frac{1}{2}$

We see that Markov's inequality is in general not very tight for $p = \frac{1}{2}$. Chebyshev's inequality performs poorly for small α. That Chebyshev's inequality is performing poorly, is not overly suprising, as the variance Bernoulli random variances is maximal when $p = \frac{1}{2}$. Hoeffding's inequality is the sharpest for all levels of α, and is very sharp when α is big.

We can explicitly compute $\mathbb{P}(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha)$, for $\alpha = 1$ and $\alpha = 0.95$. The sum of n i.i.d. Bernoulli random variables with bias p, is distributed as a binomial random variable with size paramter n and probability parameter p. Let $Z \sim \text{binom}(n, p)$. We have that

$$\mathbb{P}(\frac{1}{20}\sum_{i=1}^{20} X_i \geq 0.95) = \mathbb{P}(Y \geq 0.95) = \mathbb{P}(Z \geq 19) = \mathbb{P}(Z = 19) + \mathbb{P}(Z = 20)$$

We have an explicit formula for calculating binomial probabilities, given by

$$\mathbb{P}(Z = z) = \binom{n}{z} p^z (1-p)^{n-z}$$

Plugging in $z = 19$, $z = 20$ and $p = \frac{1}{2}$, we get that

$$\mathbb{P}(Z = 19) + \mathbb{P}(Z = 20) = 20\left(\frac{1}{2}\right)^{19}\frac{1}{2} + \left(\frac{1}{2}\right)^{20} = \frac{21}{2^{20}} = 0.00002002716 = \mathbb{P}(\frac{1}{20}\sum_{i=1}^{20} X_i \geq 0.95)$$

We also have that for $\alpha = 1$,

$$\mathbb{P}(\frac{1}{20}\sum_{i=1}^{20} X_i \geq 1) = \mathbb{P}(Z = 20) = \left(\frac{1}{2}\right)^{20} = \frac{1}{2^{20}} = 9.53674316 \cdot 10^{-7}$$

1.b

We can completely reuse the calculations from the last part of the exercise this time plugging in $p = 0.1$. We start by computing the exact probabilities $\mathbb{P}(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha)$, for $\alpha = 1$ and $\alpha = 0.95$ and $p = 0.1$. This time, we get that

$$\mathbb{P}(\frac{1}{20}\sum_{i=1}^{2} 0X_i \geq 0.95) = \mathbb{P}(Z = 19) + \mathbb{P}(Z = 20) = 20\,(0.1)^{19}\,0.9 + (0.1)^{20} = 1.81 \cdot 10^{-18}$$

We also have that

$$\mathbb{P}(\frac{1}{20}\sum_{i=1}^{20} X_i \geq 1) = \mathbb{P}(Z = 20) = \frac{1}{10^{20}}$$

We plot the empirical frequency of observing $\mathbb{P}(\frac{1}{20}\sum_{i=1}^{2} 0X_i \geq \alpha)$ as well as Markov's, Chebyshev's and Hoeffding's bound for this probability for $\alpha \in \{0.1, \cdots, 1\}$.

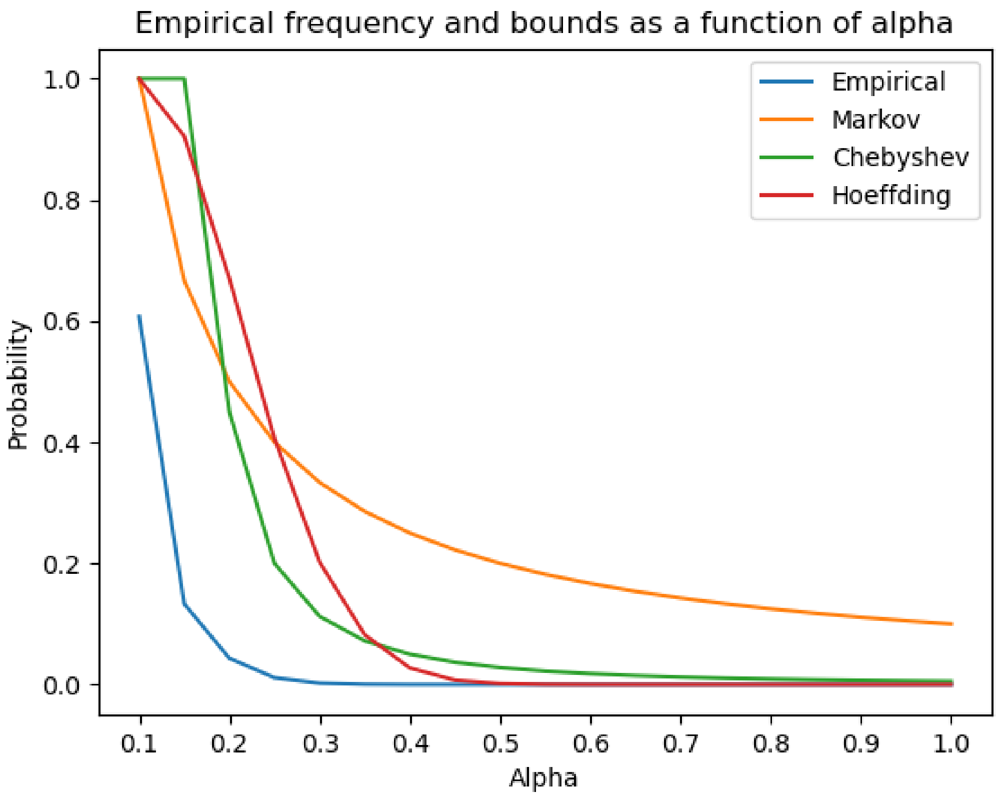

Figure 2: $p = 0.1$

We see that Markov's bound is actually the tightest for small α, as $E(X_i) = 0.1$, is small. For α around 0.25 to 0.35 Chebyshev's bound is sharpest, and for larger α Hoeffding's bound performs best.

1.c Discussion

Generally, for large α Hoeffdings inequality performance best, regardless if data is generated by $p = \frac{1}{2}$ or $p = \frac{1}{10}$. Markov's scale's poorly for large α, in both cases, but actually provides the tightest bound for a combination of small α and small p. Chebyshev's inequlity is tight when $|0.5 - p|$ is large, as the variance of X_i in this case is small - this works best for α of moderate size.

2 The Role of Independence (14 points)

Define $(X_i)_{i\in\{1,\cdots,n\}}$ such that $X_1 = Y$ where Y is a Bernoulli random variable with bias $\frac{1}{2}$. For $i \neq 1$ define X_i such that $\mathbb{P}(X_i = X_1) = 1$. We have that $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \frac{1}{2}$ for all i, and thus all the $(X_i)'$s are identically distrubted. Let $\mu = \mathbb{E}(X_i) = \frac{1}{2}$. Note that the X_i's are not independent, as we have that

$$\mathbb{P}(X_i = 0, X_j = 1) = 0 \neq \mathbb{P}(X_i = 0)\mathbb{P}(X_j = 1) = \frac{1}{4}$$

Now notice that

$$\mathbb{P}(X_1 = X_2 = \cdots = X_n = 1) = \frac{1}{2} = \mathbb{P}(X_1 = X_2 = \cdots = X_n = 0)$$

In the case $X_1 = X_2 = \cdots = X_n = 1$, we have that

$$|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i| = |\frac{1}{2} - \frac{n}{n}| = \frac{1}{2}$$

Otherwise we have that $X_1 = X_2 = \cdots = X_n = 0$, and thus

$$|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i| = |\frac{1}{2} - 0| = \frac{1}{2}$$

We thus have

$$\mathbb{P}(|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i| \geq \frac{1}{2}) = 1$$

3 Tightness of Markov's Inequality (14 points)

Let $\epsilon^* > 0$ be fixed. Define the random variable X by $\mathbb{P}(X = \epsilon^*) = \frac{1}{2} = \mathbb{P}(X = 0)$. We have that $\mathbb{E}(X) = \frac{1}{2}(0 + \epsilon^*) = \frac{\epsilon^*}{2}$. Seeing as $\epsilon^* > 0$, we also have that.

$$\mathbb{P}(X \geq \epsilon^*) = \mathbb{P}(X = \epsilon^*) = \frac{1}{2}$$

All in all we have that

$$\mathbb{P}(X \geq \epsilon^*) = \frac{1}{2} = \frac{\epsilon^*}{2\epsilon^*} = \frac{\mathbb{E}(X)}{\epsilon^*}$$

Which yields the desired result.

4 The effect of scale (range) and normalization of random variables in Hoeffding's inequality (14 points)

Let $(X_i)_{i\in\{1,\cdots,n\}}$ be independent random variables such that $P(X_i \in [0,1]) = 1$ and $E(X_i) = \mu$ for all i. Let $a_i = 0$ and $b_i = 1$ for all $i = 1, \cdots, n$. Clearly $a_i \leq b_i$ and $\mathbb{P}(X_i \in [a_i, b_i]) = 1$ for all i. Let $\epsilon > 0$ be given. Define $\epsilon^* = n\epsilon$. As $n > 0$ we have that $\epsilon^* > 0$. Thus, it follows from theorem 2.3 (2.1) that

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \geq \epsilon^*\right) \leq e^{-2(\epsilon^*)^2/\sum_{i=1}^{n}(1-0)} = e^{-2\epsilon^2 n^2/n} = e^{-2n\epsilon^2}$$

On the other hand, by the positivity of n we also have that

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \geq \epsilon^*\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \geq \epsilon\right)$$

Per the linearity of expectation we have

$$\frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

Thus,

$$= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \geq \epsilon\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right)$$

Piecing everything together, we get that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right) = \mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \geq \epsilon^*\right) \leq e^{-2n\epsilon^2}$$

Which yields the desired result.

5 Logistic regression

5.1 Cross-entropy error measure (11 points)

(a)

Assume we are learning from ±1 i.i.d. data to predict a noisy target

$$\mathbb{P}(y|x) = \begin{cases} f(x), & y = 1, \\ 1 - f(x), & y = -1 \end{cases}$$

Given a hypothesis $h(x)$ in the hypothesis set the likelihood would then be

$$\mathbb{P}(y|x) = \begin{cases} h(x), & y = 1, \\ 1 - h(x), & y = -1 \end{cases}$$

As the logarithm is monotonically increasing maximizing the likelihood is equivalent to maximizing the log-likelihood. As $x \mapsto -x$ is monotonically decreasing, this is again equivalent to minimizing the negative log-likelihood, that is, minimizing

$$\ell(h) = -\log\left(\prod_{n=1}^{N} \mathbb{P}(y_n|x_n)\right) = \sum_{n=1}^{N} -\log(\mathbb{P}(y_n|x_n)) = \sum_{n=1}^{N} \log\frac{1}{\mathbb{P}(y_n|x_n)}$$

Notice, for a hypothesis h, we have that

$$\mathbb{P}(y|x) = \mathbb{1}(y = 1)h(x) + \mathbb{1}(y = -1)(1 - h(x))$$

Thus we have that

$$\ell(h) = \sum_{n=1}^{N} \log\frac{1}{\mathbb{1}(y_n = 1)h(x_n) + \mathbb{1}(y_n = -1)(1 - h(x_n))}$$

Notice that when $y_n = 1$, the corresponding term in the series is equal to $\log\frac{1}{h(x_n)}$, and when $y_n = -1$ the corresponding term in the series is equal to $\log\frac{1}{1-h(x_n)}$. The task of maximizing the likelihood therefore boils down to minimizing

$$E_{\text{in}}(h(x)) = \sum_{n=1}^{N} \mathbb{1}(y_n = 1)\log\frac{1}{h(x_n)} + \mathbb{1}(y_n = -1)\log\frac{1}{1 - h(x_n)}$$

(b)

It is sufficient to show that

$$E_{\text{in}}(h(x)) = \sum_{n=1}^{N} \mathbb{1}(y_n = 1) \log \frac{1}{h(x_n)} + \mathbb{1}(y_n = -1) \log \frac{1}{1 - h(x_n)} = \sum_{n=1}^{N} \log(1 + \exp(-y_n w^T x_n))$$

As multiplying the expression in 3.9 by N, does not change it's minimum. To show the above equality it is sufficient to show that each term in the respective sums are equal. Set $h(x) = \theta(w^T x)$, where $\theta(s) = \frac{e^s}{1+e^s}$. Note that

$$\theta(-s) = \frac{e^{-s}}{1 + e^{-s}} = \frac{1}{e^s + 1} = 1 - \frac{e^s}{1 + e^s} = 1 - \theta(s)$$

Assume that $y = 1$. Then

$$\mathbb{1}_{y=1}(1) \log \frac{1}{h(x)} + \mathbb{1}_{y=-1}(1) \log \frac{1}{1 - h(x)} = \log \frac{1}{\theta(w^T x)} = \log \left(\frac{1 + e^{w^T x}}{e^{w^T x}} \right) = \log \left(\frac{1 + e^{-w^T x}}{1} \right) = \log \left(1 + e^{-y w^T x} \right)$$

Now consider the case $y = -1$

$$\mathbb{1}_{y=1}(-1) \log \frac{1}{h(x)} + \mathbb{1}_{y=-1}(-1) \log \frac{1}{1 - h(x)} = \log \frac{1}{1 - \theta(w^T x)} = \log \frac{1}{\theta(-w^T x)} = \log(1 + e^{w^T x}) = \log(1 + e^{-y w^T x})$$

All in all we have that

$$\log(1 + e^{-y w^T x}) = \mathbb{1}(y = 1) \log \frac{1}{h(x)} + \mathbb{1}(y = -1) \log \frac{1}{1 - h(x)}$$

For all y, x, w, which yields the desired result.

5.2 Logistic regression loss gradient (13 points)

The case with $y = \pm 1$

We have that $(ln(x))' = \frac{1}{x}$, and that by the chain rule

$$\frac{\partial(1 + \exp(-y w^T x))}{\partial w^T} = -yx \exp(-y w^T x)$$

Again, by the chain rule, we have that

$$\frac{\partial \log((1 + \exp(-y w^T x))}{\partial w^T} = \frac{1}{1 + \exp(-y w^T x)} - yx \exp(-y w^T x) = \frac{-yx}{1 + \exp(y w^T x)}$$

By definiton of as we have seen previously, $\theta(-s) = \frac{1}{e^s + 1}$, and we can then also write

$$\frac{-yx}{1 + \exp(y w^T x)} = -yx\theta(-y w^T x)$$

As the derivative is a linear operator, we have that

$$\nabla E_{\text{in}}(w) = \nabla \frac{1}{N} \sum_{n=1}^{N} \log(1 + \exp(-y_n w^T x_n)) = -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n x_n}{1 + \exp(y_n w^T x_n)} = \frac{1}{N} \sum_{n=1}^{N} -y_n x_n \theta(-y_n w^T x_n)$$

We have that the absolute size of the gradient depends on

$$\theta(-yw^T x) = \frac{1}{1 + e^{yw^T x}}$$

This term is large when, $|yw^T x| >> 0$ and $yw^T x < 0$. The second condition captures a misclassification, as $yw^T x$ is negative if and only if $\text{sign}(y) \neq \text{sign}(w^T x)$. That is, terms where a misclassification has happened, contributes more to the gradient than when a label is correctly classified. Further terms where a severe misclassification has happened, that is the algorithm guessing the wrong label with a high likelihood contributes more to the gradient, as this is equivalent to the first condition.

The case with $y \in \{0, 1\}$

If we assume that the labels are $y \in \{0, 1\}$, we want to show that the gradient of the negative log-likelihood is given as

$$-\frac{1}{N}\sum_{n=1}^{N}(y_n - \theta(w^T x_n))x_n$$

We utilize that we have already covered the case for $y = \pm 1$. In this setting, assume that $y = 1$, then

$$\frac{yx}{1 + e^{yw^T x}} = \frac{x}{1 + e^{w^T x}} = \theta(-w^T x)x = (1 - \theta(x))x = \left(\frac{y+1}{2} - \theta(w^T x)\right)x$$

Assume now that $y = -1$, then

$$\frac{yx}{1 + e^{yw^T x}} = \frac{-x}{1 + e^{-w^T x}} = -\theta(w^T x)x = \left(\frac{y+1}{2} - \theta(w^T x)\right)x$$

Therefore, all in all, we can conlude that

$$-\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1 + \exp(y_n w^T x_n)} = -\frac{1}{N}\sum_{n=1}^{N}\left(\frac{y_n + 1}{2} - \theta(w^T x)\right)x_n$$

Consider the mapping $y \mapsto \frac{y+1}{2}$. When $y = 1$, this is equal to 1, and when $y = -1$, this is equal to 0. That is, this mapping transforms the labels from $y = \pm 1$ to $y \in \{0, 1\}$. We can simply plug these transformed labels into our gradient from before, to arrive at the result, that for $y \in \{0, 1\}$ we have that

$$\nabla E_{\text{in}}(w) = -\frac{1}{N}\sum_{n=1}^{N}(y_n - \theta(w^T x_n))x_n$$

Notice that we can write

$$y - \theta(w^T x) = \begin{cases} -\theta(w^T x), & y = 0 \\ \theta(-w^T x), & y = 1 \end{cases}$$

Recall that in this parametrization, the linear part of the model encodes log-odds, that is

$$w^T x = \ln \frac{\mathbb{P}(y = 1|X = x)}{\mathbb{P}(y = 0|X = x)}$$

Combining this, with our piecewise expression for $y - \theta(w^T x)$, we see that once again the contribution to the gradient is greatest from the misclassified cases.

5.3 Log-odds (11 points)

Consider binary logistic regression. Let the input space be $\mathbb{R}^d$, the label space $\{0,1\}$. Define the model f with parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ as

$$f(x) = \sigma(w^T x + b) = \mathbb{P}(Y = 1|X = x)$$

Assume that

$$w^T x + b = \log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)}$$

Let $s = w^T x + b$. We want to show that σ is the logistic function, i.e. that $\sigma(s) = \frac{1}{1+e^{-s}}$. As $\mathbb{P}$ is a probability measure, we have that

$$s = \log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} \Leftrightarrow \exp(s)(1 - \mathbb{P}(Y = 1|X = x)) = \mathbb{P}(Y = 1|X = x)$$

We also have that $\mathbb{P}(Y = 1|X = x) = \sigma(s)$, which leads us to

$$\exp(s)(1 - \sigma(s)) = \sigma(s) \Leftrightarrow \exp(-s) = \frac{\sigma(s)}{1 - \sigma(s)} \Leftrightarrow \exp(-s) + 1 = \frac{1}{\sigma(s)} = \Leftrightarrow \frac{1}{1 + \exp(-s)} = \sigma(s)$$

Which is what we wanted to show.