Machine Learning A - Home Assignment 2

Jakob H. Schauser, pwn274

September 2021

1 Illustration of Markov's, Chebyshev's, and Hoeffding's Inequalities

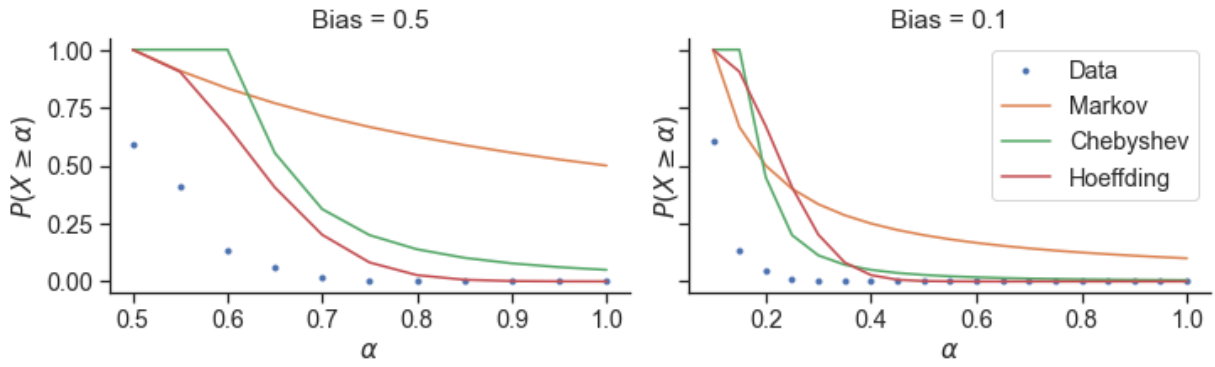

Figure 1: The two experiments. The simulation of flipping 20 coins a million times. In the second experiment, the coin was biased.

To illustrate the three inequalities, a simulation was made. The code relied mainly on Numpy's random library. The results can be seen in figure 1.

The granularity of the x-axis comes from the fact that we are only flipping 20 coins at a time. This means the percentage of heads can only jump by 5% i.e. α moves in steps of 0.05.

Exactly as I had expected, the Hoeffding inequality is the tightest, followed by Chebyshev and lastly the Markov inequality has the worst concentration. This also mostly holds for the system with bias 0.1 - especially in the low-probability limit,

Calculating the exact probabilities is pretty easy, as 1 and 0.95 simply corresponds to flipping "all heads" or "all heads except one". This corresponds to $1/2^{20}$ and $20/2^{20}$ respectively.

2 The Role of Independence

After running my head into the wall, multiple times I got a hint to look for the simplest system where all coins were dependent on the first one. The simplest system I could think up: Given the first flipped coin, flip all other coins to the same. I.e. if the first coin gives 1, the $n-1$ next coins are all 1. The expectation value for any single coin is still 0.5, but the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ will be either 0 or 1 depending on the first - either of which gives $\left|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i\right| = 0.5$, meaning the probability is:

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

Exactly as it was supposed to be!

3 Tightness of Markov's Inequality

I used the hint and looked at random variables that could take one of two possible values. I found, that a random variable with a fifty-fifty chance of being either 0 or ϵ^* makes the inequality hold as an equality. This can be seen as:

$$\mathbb{P}(X \geq \varepsilon^*) = 0.5 = \frac{\mathbb{E}[X]}{\varepsilon^*} = \frac{0.5 \cdot \varepsilon^*}{\varepsilon^*} = 0.5 \tag{1}$$

Here I have used that $\mathbb{P}(X \geq \varepsilon^*) = 0.5$ by design and $\mathbb{E}[X] = 0.5\varepsilon^*$ which can easily be verified as there are only two outcomes.

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

Starting from the original definition of Hoeffding's inequality:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2} \tag{2}$$

The first point is realizing, that this is correct for any $a_i < b_i$ in our domain, allowing us to choose the extremes (0 and 1 respectively) thereby giving us:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(1-0)^2} = e^{-2\varepsilon^2/n} \tag{3}$$

Next, you see that the $X_1, .., X_i$ are identical distributions, allowing the following rewrite:

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = n\mu \tag{4}$$

Combining what we have so far gives:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - n\mu \geq \varepsilon\right) \leq e^{-2\varepsilon^2/n} \tag{5}$$

The next-to-final step now consists of dividing both sides of the inequality inside the probability by n:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \frac{1}{n}\varepsilon\right) \leq e^{-2\varepsilon^2/n} \tag{6}$$

Now as we can freely choose ε, we define $\varepsilon' = \frac{1}{n}\varepsilon$, giving us:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon'\right) \leq e^{-2n\varepsilon'^2} \tag{7}$$

The answer we set out to prove!

5 Preprocessing & Regularization

a)

I know that $\text{var}(x_1) = \text{var}(\hat{x}_1) = 1$

Finding the variance of x_2 is slightly harder, but doable as follows:

$$\begin{aligned}
\text{var}(x_2) &= \text{var}\left(\sqrt{1-\varepsilon^2}\hat{x}_1 + \varepsilon\hat{x}_2\right) \\
&= \text{var}\left(\sqrt{1-\varepsilon^2}\hat{x}_1\right) + \text{var}\left(\varepsilon\hat{x}_2\right) \\
&= \left(1-\varepsilon^2\right)\text{var}(\hat{x}_1) + \varepsilon^2\text{var}(\hat{x}_2) \\
&= \left(1-\varepsilon^2\right) + \varepsilon^2 \\
&= 1
\end{aligned}$$

Where I have used the independence of $\hat{x}_1$ and $\hat{x}_2$ and the fact that $\text{var}(aX) = a^2\text{var}(X)$

The covariance is found by simply inserting the variables in the definition;

$$
\begin{aligned}
cov(x_1, x_2) &= \mathbb{E}[(x_1 - \mathbb{E}[x_1])(x_2 - \mathbb{E}[x_2])] \\
&= \mathbb{E}[(\hat{x}_1 - \mathbb{E}[\hat{x}_1])(\sqrt{1-\varepsilon^2}\hat{x}_1 + \varepsilon\hat{x}_2 - \mathbb{E}[\sqrt{1-\varepsilon^2}\hat{x}_1 + \varepsilon\hat{x}_2])] \\
&= \mathbb{E}[(\hat{x}_1 - 0)(\sqrt{1-\varepsilon^2}\hat{x}_1 + \varepsilon\hat{x}_2 - 0)] \\
&= \mathbb{E}[\sqrt{1-\varepsilon^2}\hat{x}_1^2 + \varepsilon\hat{x}_2\hat{x}_1] \\
&= \mathbb{E}[\sqrt{1-\varepsilon^2}\hat{x}_1^2] + \mathbb{E}[\varepsilon\hat{x}_2\hat{x}_1] \\
&= \sqrt{1-\varepsilon^2}\mathbb{E}[\hat{x}_1^2] + \varepsilon\mathbb{E}[\hat{x}_2\hat{x}_1] \\
&= \sqrt{1-\varepsilon^2}\mathbb{E}[\hat{x}_1^2] + \varepsilon\mathbb{E}[\hat{x}_2]\mathbb{E}[\hat{x}_1] \\
&= \sqrt{1-\varepsilon^2}(\mathbb{E}[\hat{x}_1]^2 + \mathrm{var}(\hat{x}_1)) + 0 \\
&= \sqrt{1-\varepsilon^2}(0+1) \\
&= \sqrt{1-\varepsilon^2}
\end{aligned}
$$

b)

Using the definitions for x_1 and x_2:

$$
\begin{aligned}
f(x) &= \hat{\omega}_1\hat{x}_1 + \hat{\omega}_2\hat{x}_2 \\
&= \hat{\omega}_1 x_1 + \hat{\omega}_2(x_2 - \sqrt{1-\varepsilon^2}\hat{x}_1)/\varepsilon \\
&= (\hat{\omega}_1 - \hat{\omega}_2\sqrt{1-\varepsilon^2}/\varepsilon)x_1 + (\hat{\omega}_2/\varepsilon)x_2
\end{aligned}
$$

Here, f is clearly linear in x_1 and x_2 with $\omega_1 = \hat{\omega}_1 - \hat{\omega}_2\sqrt{1-\varepsilon^2}/\varepsilon$ and $\omega_2 = \hat{\omega}_2/\varepsilon$

c)

The regularization constraint gives:

$$
\begin{aligned}
C \geq \omega_1^2 + \omega_2^2 &= (\hat{\omega}_1 - \hat{\omega}_2\sqrt{1-\varepsilon^2}/\varepsilon)^2 + (\hat{\omega}_2/\varepsilon)^2 \\
&= \hat{\omega}_1^2 + \hat{\omega}_2^2(1-\varepsilon^2)/\varepsilon^2 - 2\hat{\omega}_1\hat{\omega}_2\sqrt{1-\varepsilon^2}/\varepsilon + \hat{\omega}_2^2/\varepsilon^2 \\
&= 1 + (1-\varepsilon^2)/\varepsilon^2 - 2\sqrt{1-\varepsilon^2}/\varepsilon + 1/\varepsilon^2 \\
&= 2\left(1/\varepsilon^2 - \sqrt{1-\varepsilon^2}\right)/\varepsilon
\end{aligned}
$$

d)

As $\varepsilon \to 0$ the first term dominates and goes towards infinity as ε^2. Meaning the lower bound on C $\to \infty$ when the correlation rises. This can be seen visually on the plot in figure 2.

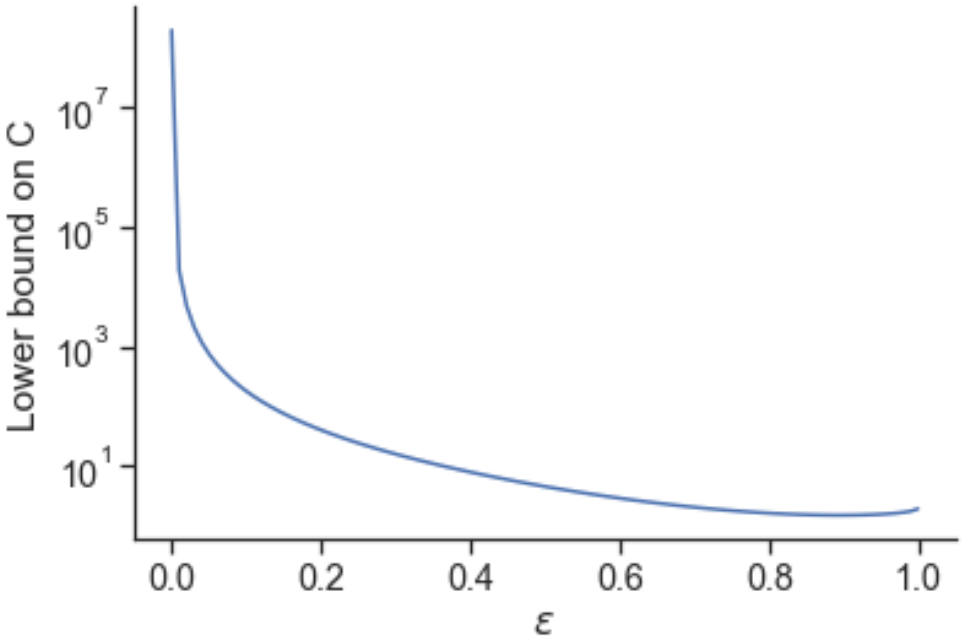

Figure 2: Caption