Machine Learning - First Hand-In

Lucas Johan Boesen - kzx651

November 30, 2021

1 Make Your Own

1.1 Profile Information

Lets assume we are to do this in Denmark. I would gather the following for each available student:
The students given grades from all courses $x \in \{-3, 0, 2, 4, 7, 10, 12\}$
The given semester in which the student joined a given course, $x \in \{1, 2, 3, 4, 5, 6.7, 8, 9, 10\}$
Sex $x \in \{F, M\}$
Age $x \in \{18, ...\}$
From the data regarding the courses I could do some feature engineering and calculate no. of courses in statistics, linear algebra etc, and use them as well. I would collect these observations in a Model Matrix, where I would one hot encode the categorical and keep the numeric ones as they are (unless any transformation is deemed necessary).

1.2 Label Space

The label space would have to be the possible grades that one can achieve, and thereby $Y \in \{-3, 0, 2, 4, 7, 10, 12\}$, which I would collect for all the students available.

1.3 Loss Function

I would use squared loss so that we can punish wrong estimates that are further away from the actual observed harder than the ones that are close, e.g. if we guess that student A gets 12 and he/she actually gets 4, then it would be worse for the student compared to guessing 7 and he/she gets 4.

1.4 Distance Measure

I would use the Eucledian Distance for my distance measure.

1.5 Performance of the algorithm

To evaluate the performance, I would try to minimize the loss function using cross-validation or train-test splits.

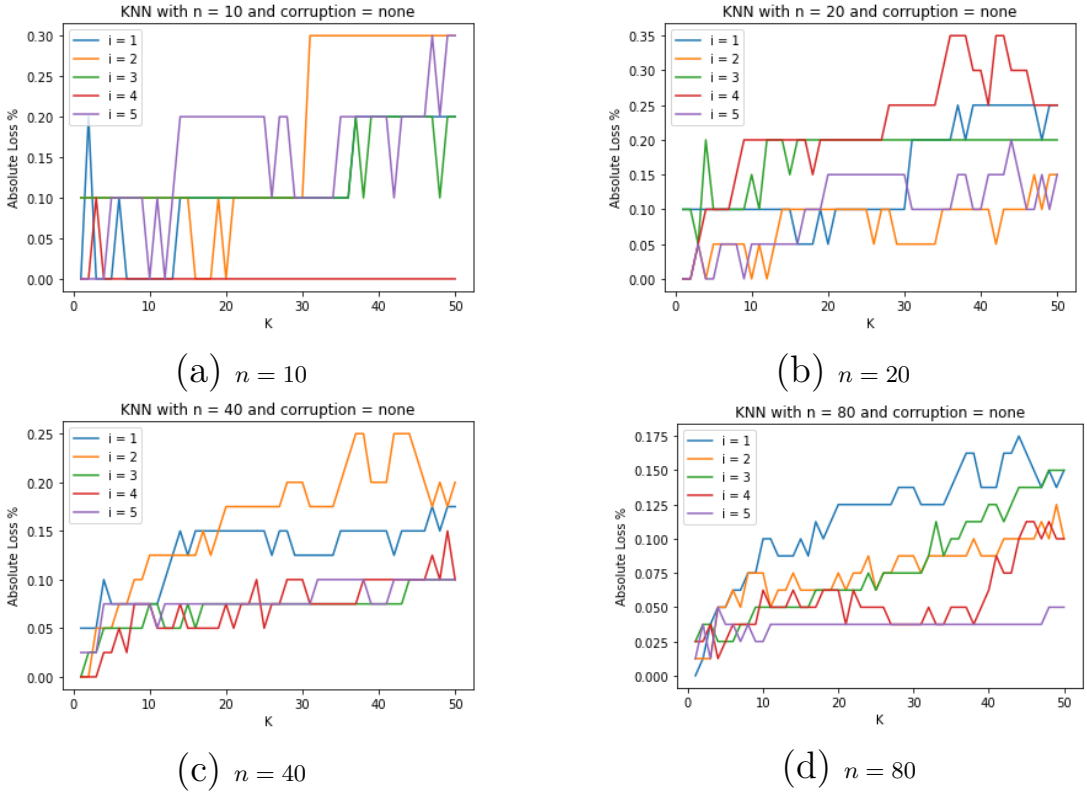
1.6 Alleviation of issues

If we get perfect performance I could imagine that we overfitted when minimizing the loss function, and would thereby try other values of K, train size and so on, if I where to use other algorithms than KNN I would also consider some form of regularization; lasso, ridge...

2 Digits Classification with K Nearest Neighbors

2.1 Tast #1

For Loss function we have just implemented the Absolute Loss even though it does not really matter whether we use Zero-one loss, squared loss or absolute loss in this situation.
Looking at Figure 1 we can see that as we increase n we decrease our loss, moreover the fluctuations of errors are also more wild when n is lower. Regarding K we can see that as K increases our loss increases. The prediction accuracy is thereby getting worse when we increase K, all though the best K might not be exactly 1 but generally speaking as K increases our prediction error gets worse. The optimal K is very low and probably between 1-4.
The variance of the validation error can be seen in Figure 2 and it shows that as K increases our variance of the validation error also increases.



(a) $n = 10$ (b) $n = 20$

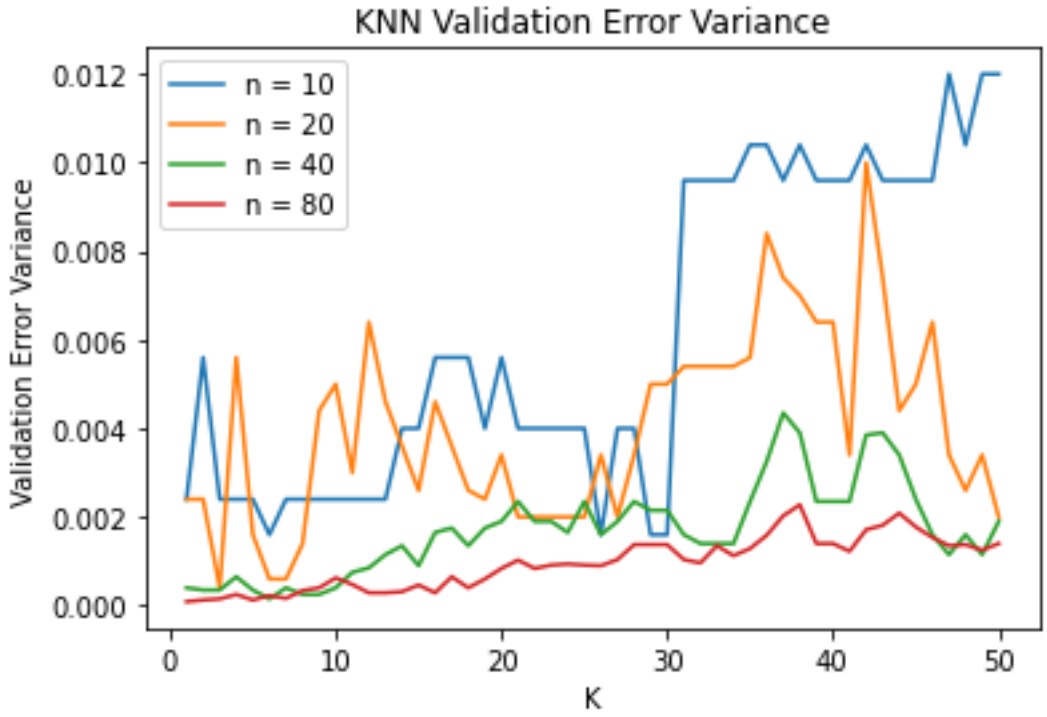
(c) $n = 40$ (d) $n = 80$

Figur 1: plots of validation error for KNN with $K \in \{1, ..., 50\}$ and $n \in \{10, 20, 40, 80\}$ for five different sets of size n.

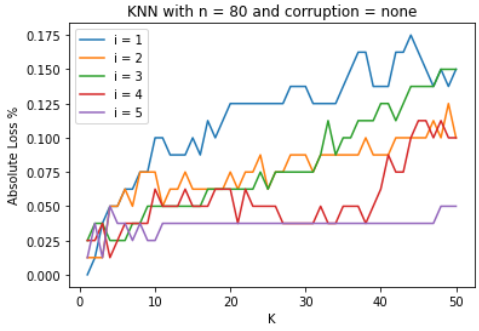
To calculate the variance we have simply taken the variance for the 5 observations for each of the tested $K's$ and plotted as shown in Figure 2.

2.2 # Tast 2

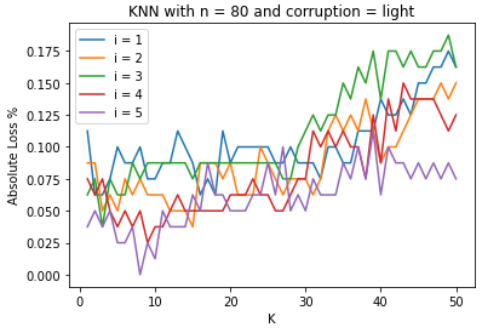
The corrupted datasets shows a clear correlation with the error and the amount of corruption; obviously having heavy corrupted as the worst performing and the light corrupted the best out of the corrupted ones. We can also see that the optimal value of K is now larger than it was without corruption, and is probably between 5-10. This can be seen in Figure 3.



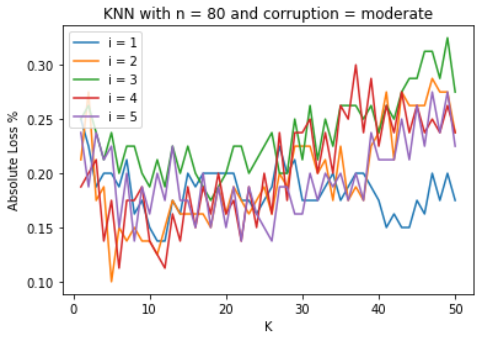
Figur 2: Variance for validation errors over the five validation sets



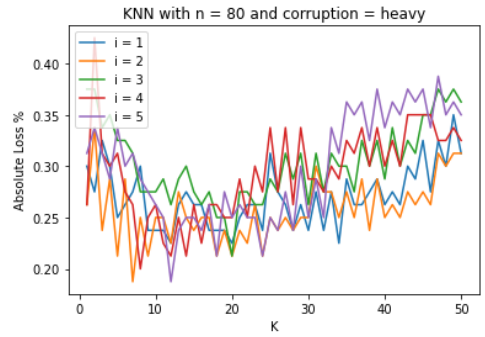
(a) None Corruption



(b) Light Corruption



(c) Moderate Corruption

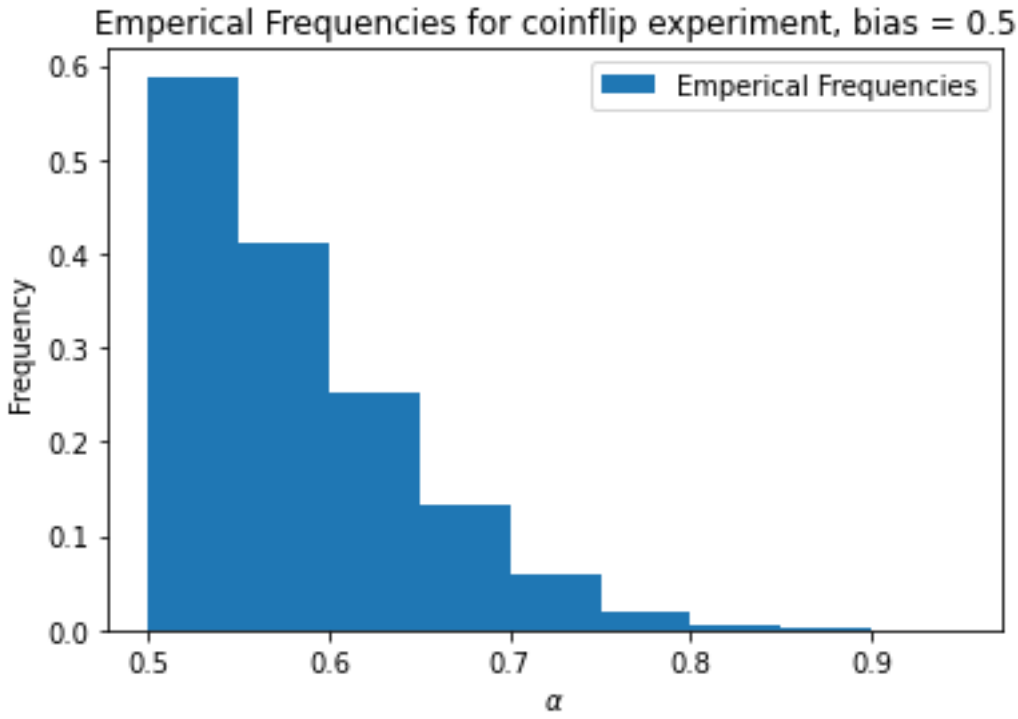


(d) Heavy Corruption

Figur 3: plots of validation error for KNN with $K \in \{1, ..., 50\}$ and $n \in \{10, 20, 40, 80\}$ for five different sets of size n.

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

3.1 Emperical Frequency



3.2 Explanation

Since we only have 20 observations for each of the repetitions, our observation space for the experiment is

$$\text{flips} \in \{\frac{1}{20}, \frac{2}{20}, ..., 1\}$$

therefore a granularity of $\alpha = 0.51$ is equal to $\alpha = 0.55$.

3.3 Markov's Bound

Markov's Bound is as follows

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$$

where we will use $\epsilon = \alpha$. The bound in our situations is as follows

$$\mathbb{P}\left(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\right) \leq \frac{\mathbb{E}\left[\frac{1}{20}\sum_{i=1}^{20} X_i\right]}{\alpha}$$

Lets denote a heads flip as 1 and a tails flip as 0, and since $\alpha = 0.5$, the expected value for a single flip is

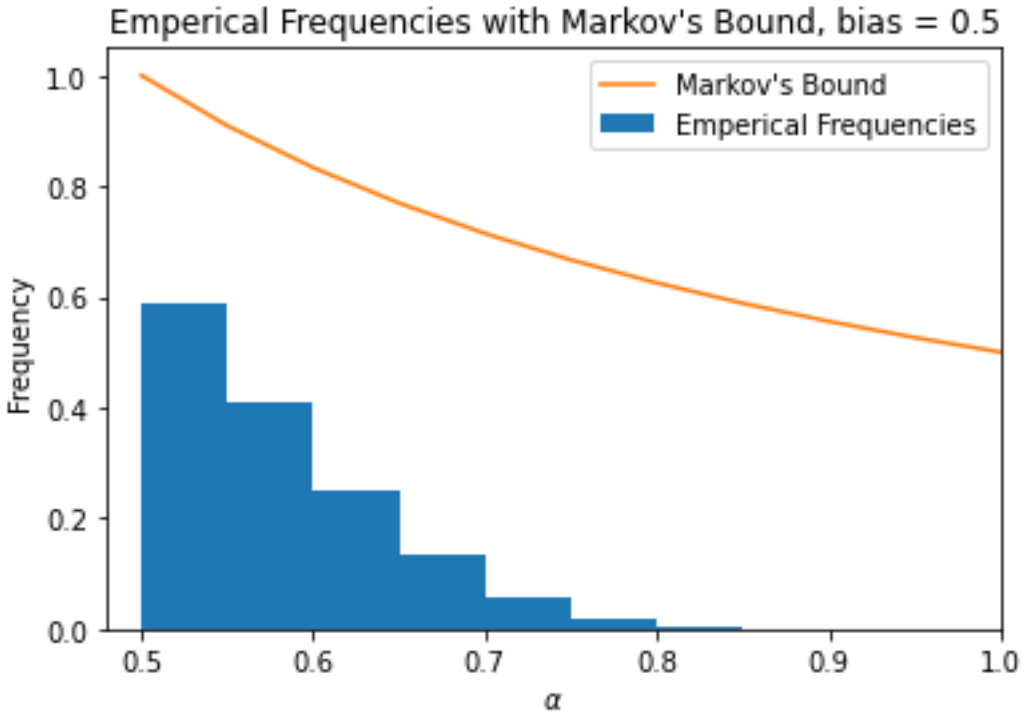
$$\mathbb{E}[X_i] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2}$$

and thereby

$$\mathbb{E}\left[\frac{1}{20}\sum_{i=1}^{20} X_i\right] = \frac{1}{2}$$

which leads to the RHS as

$$\frac{\mathbb{E}[X]}{\alpha} = \frac{1}{2\alpha}$$



3.4 Chebyshev's Inequality

Chebyshev's inequality is as follows:

$$\mathbb{P}(|X - \mathbb{E}[X]|) \geq \epsilon) \leq \frac{\text{Var}\,[X]}{\epsilon^2}$$

where X is the same as earlier. By using Markov's Inequality as shown in the lectures, and can thereby define Chebyshev's Bound

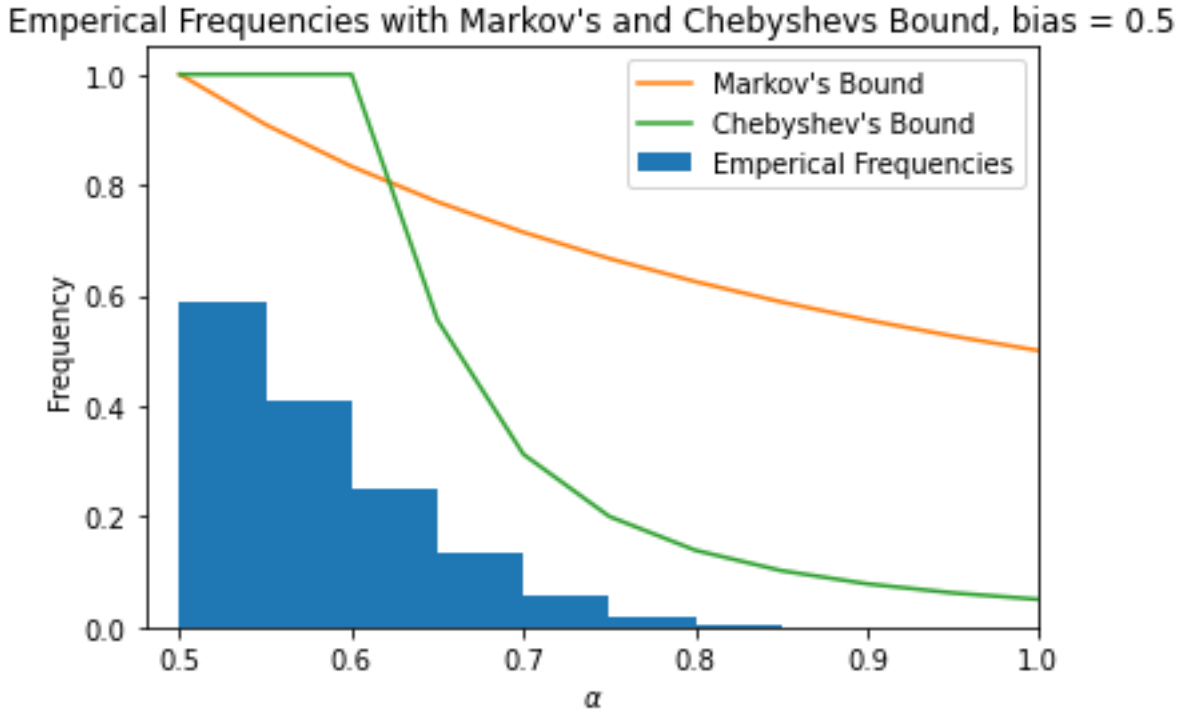
$$\mathbb{P}(\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n] \geq \epsilon) \leq \mathbb{P}(|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n]| \geq \epsilon) \leq \frac{\text{Var}\,[\hat{\mu}_n]}{\epsilon^2} = \frac{\text{Var}\,[X_1]}{n\epsilon^2}$$

and still $\mathbb{E}[X_1] = \frac{1}{2}$. By calculating $\text{Var}\,[\hat{\mu}_n]$

$$\begin{aligned}
\text{Var}\,[X] &= p - p^2 \\
\text{Var}\,[\hat{\mu}_n] &= \text{Var}\left[\frac{1}{20}\sum_{i=1}^{20} X_i\right] \\
&= \left(\frac{1}{20}\right)\text{Var}\,[X] \\
&= \frac{1}{20}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{80}
\end{aligned}$$

using this we get the bound

$$\mathbb{P}(X \geq \alpha) = \mathbb{P}(X - \mathbb{E}[X] \geq \alpha - \mathbb{E}[X]) \leq \frac{1}{80 \cdot (\epsilon - \frac{1}{2})^2}$$



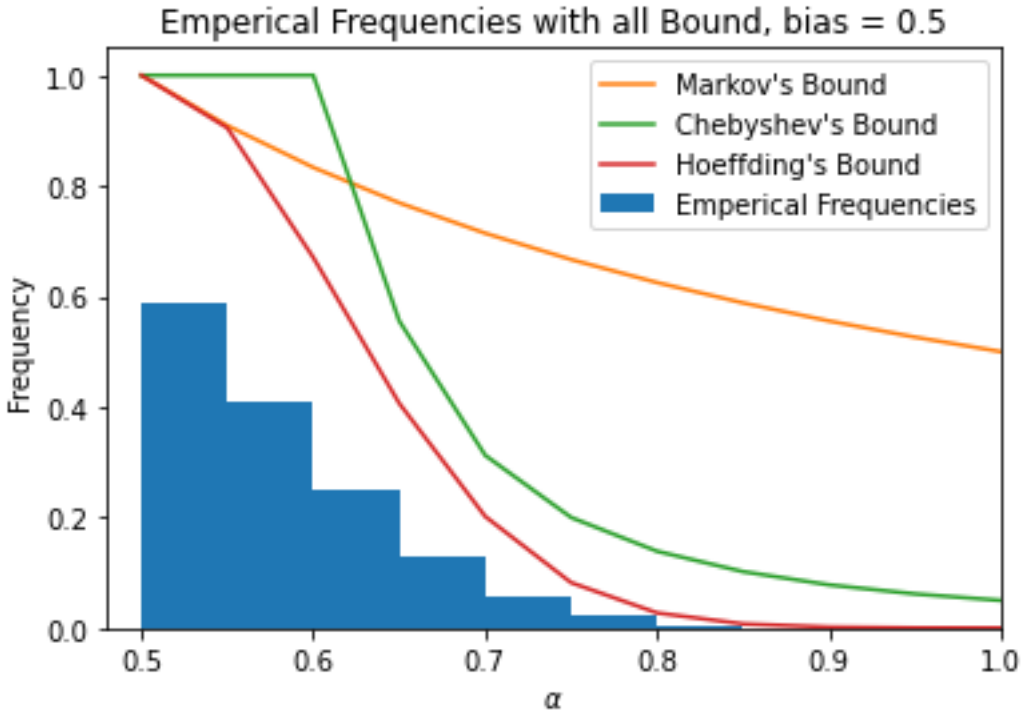
3.5 Hoeffding's Bound

We get the Hoeffding's one-sided Inequalities as

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

for every $\epsilon > 0$ and where $X_i \in [a_i, b_i]$ and in our case $a = 0$ and $b = 1$.
Using the same X as earlier we can write it on the wanted form by

$$\mathbb{P}(X - \mathbb{E}[X] \geq \alpha) = \mathbb{P}(X \geq \alpha - \mathbb{E}[X]) \leq e^{-2n(\alpha - \frac{1}{2})^2}$$



3.6 Comparison between the four plots

When asked about a comparison between the four plots, I assume that one is asked about a comparison between the bounds since this is the only thing that is changed from each plot.

We can see that Hoeffding's bound is lower than both Chebyshev's and Markov's bounds at all values for α, while Chebyshev's is worse than Markovs until $\alpha = 0.65$. This follows the theory since Hoeffding's bound is the best of the three.

3.7 Exact probability

The probability to draw at least k is as follows

$$F(k; n, p) = \mathbb{P}(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

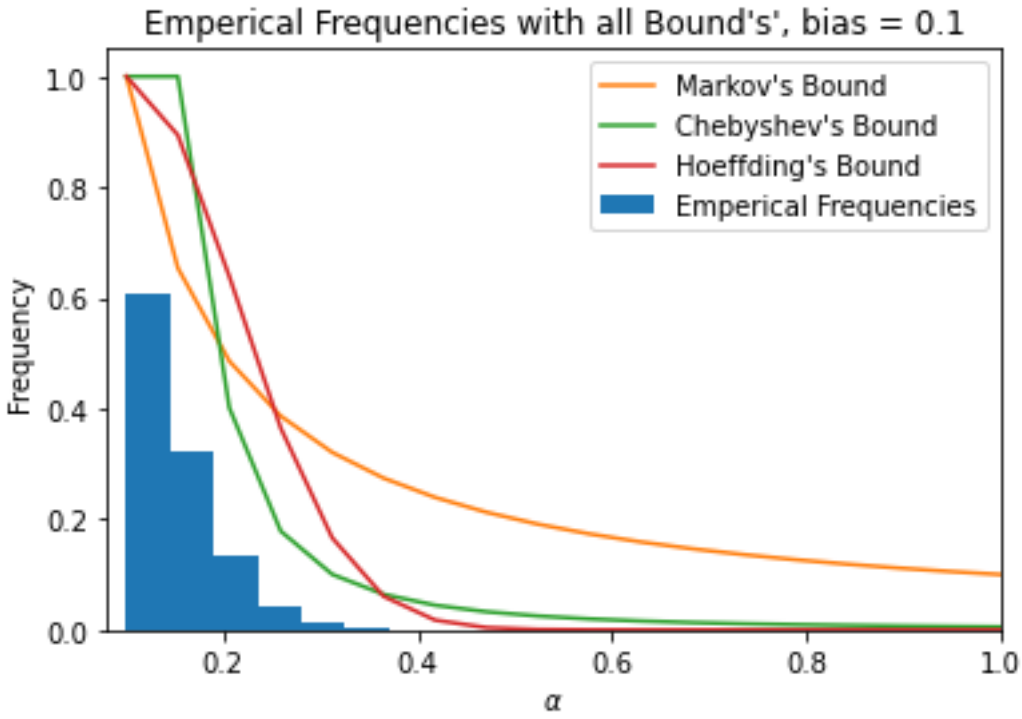
For $\alpha = 1$ it is as simple as follows (still same X as earlier...), multiplying both sides in the inequality by 20 we get

$$\mathbb{P}(20X \geq 20) = \binom{20}{20} \left(\frac{1}{2}\right)^{20} \left(1 - \frac{1}{2}\right)^{20-20} = \left(\frac{1}{2}\right)^{20}$$

for $\alpha = 0.95$ we get

$$\begin{aligned} \mathbb{P}(20X \geq 20) &= \sum_{i=19}^{20} \binom{20}{i} \frac{1}{2}^i \left(1 - \frac{1}{2}\right)^{20-i} \\ &= \left(\frac{1}{2}\right)^{20} + \left(\frac{1}{2}\right)^{19} \left(\frac{1}{2}\right)^{1} \binom{20}{19} \\ &\approx \text{2e-05} \end{aligned}$$

3.8 Repeating question with bias of 0.1



The exact probabilities follow the same logic as earlier, and is:

$$\mathbb{P}(20X \geq 20) = \left(\frac{1}{10}\right)^{20}$$

$$\mathbb{P}(20X \geq 19) = \mathbb{P}(20X \geq 20) + \left(\frac{1}{2}\right)^{19}\left(\frac{1}{2}\right)^{1}\binom{20}{19} \approx 1.8\text{e-}18$$

3.9 Discuss the results

The biggest change we get when changing the bias, is that we now can see that Chebyshev's and Hoeffding's Bounds is better than Markov's when we are more in the right tail, while Markov's bound is lower when we are around $\alpha \in \{0.1, ..0.2\}$. Since both Chebyshev's and Hoeffding's bound get more exact when increasing n I would expect them to be better at lower α's if we increased n.

4 Basic Linear Algebra

We consider the hyperplane

$$w^\top x + b = 0$$

and wants to find the distance from the hyperplane to the origin. The distance from a point, x_0, to the hyperplane is

$$d = \frac{|wx_0 + b|}{\|w\|}$$

having $x_0 = 0$ we get the distance from the hyperplane to the origin:

$$d = \frac{|b|}{\|w\|}$$

5 Regression

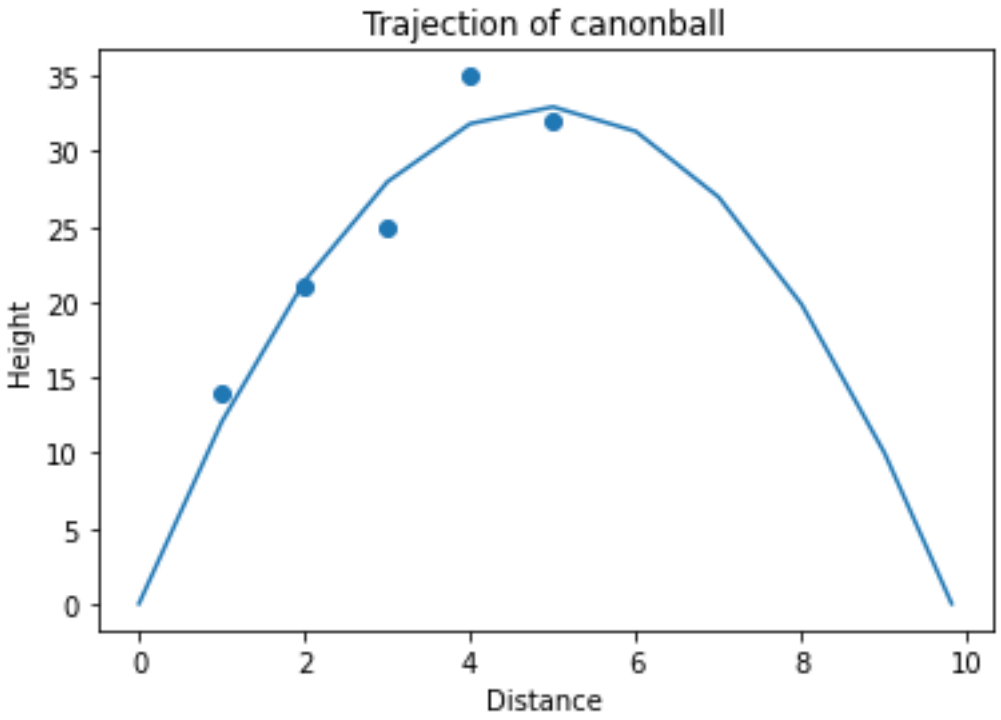
The expression describing the parabolic trajection is a 2. degree polynomium on the form

$$
\begin{aligned}
Y &= X\beta + \epsilon \\
\beta &= (X^\top X)^{-1} X^\top Y \\
X &= \begin{bmatrix} x_1 & x_1^2 \\ \vdots & \vdots \\ x_n & x_n^2 \end{bmatrix} \\
Y &= \begin{bmatrix} y_i \\ \vdots \\ y_n \end{bmatrix}
\end{aligned}
$$

In our case β is

$$
\begin{aligned}
\beta &= \begin{bmatrix} 13.43913043 \\ -1.36956522 \end{bmatrix} \\
y_i &= 13.43913043 x_i - 1.36956522 x_i^2 + \epsilon
\end{aligned}
$$

It should be noted that we omit the intercept, to allow the canon to start in (0, 0). To get the destination where the canonball lands we need to find the root which is not equal to 0 ($x \neq 0$), doing this shows that the ball lands around 9.8127, which can also be seen in Figure 4.



Figur 4: Estimated trajection of canonball using least squares on a 2. degree polynomium