# SVD

## Theory

1. Suppose $A \in \mathbb{R}^{mxn}$ represents a transformation from vector space $V \subseteq \mathbb{R}^n \, to \, W \subseteq \mathbb{R}^m$.

   (a) $im(A) = \{w \in W : w = Ax, x \in V\}, im(A) \subseteq \mathbb{R}^m$. Let's show that im(A) is a subspace of $\mathbb{R}^m$:

   i. $A\vec{0} = \vec{0} \Rightarrow im(A) \neq \emptyset$

   ii. Let $w_1, w_2 \in im(A), \lambda_1, \lambda_2 \in \mathbb{R}$
   $\exists v_1, v_2 \in V \ s.t \ Av_1 = w_1, Av_2 = w_2$

   $$\lambda_1 w_1 + \lambda_2 w_2 = \lambda_1 A v_1 + \lambda_2 A v_2 = A(\lambda_1 v_1 + \lambda_2 v_2)$$

   and since $\lambda_1 v_1 + \lambda_2 v_2 \in V$ and $A(\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1 w_1 + \lambda_2 w_2 \in W$ we get

   $$\lambda_1 w_1 + \lambda_2 w_2 \in im(A)$$

   (b) $im(A^T) = \left\{x \in V : x = A^T w, w \in W\right\}, im(A^T) \subseteq \mathbb{R}^n$. Let's show that $im(A^T)$ is a subspace of $\mathbb{R}^n$:

   i. $A^T\vec{0} = \vec{0} \Rightarrow im(A^T) \neq \emptyset$

   ii. Let $v_1, v_2 \in im(A^T), \lambda_1, \lambda_2 \in \mathbb{R}$
   $\exists w_1, w_2 \in W \ s.t \ Aw_1 = v_1, Aw_2 = v_2$

   $$\lambda_1 w_1 + \lambda_2 w_2 = \lambda_1 A v_1 + \lambda_2 A v_2 = A(\lambda_1 v_1 + \lambda_2 v_2)$$

   and since $\lambda_1 v_1 + \lambda_2 v_2 \in V$ and $A(\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1 w_1 + \lambda_2 w_2 \in W$ we get

   $$\lambda_1 w_1 + \lambda_2 w_2 \in im(A)$$

   (c) $ker\,(A) = \{x \in V : Ax = 0\} \subseteq \mathbb{R}^n$, again, let's show that $ker(A)$ is a subspace of $\mathbb{R}^n$.

   i. $A\vec{0} = \vec{0} \Rightarrow ker\,(A) \neq \emptyset$

   ii. Let $v_1, v_2 \in ker\,(A), \lambda_1, \lambda_2 \in \mathbb{R} \Rightarrow A(\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1 A v_1 + \lambda_2 A v_2 = 0 \Rightarrow \lambda_1 v_1 + \lambda_2 v_2 \in ker(A)$

2. Let T be a linear transformation $T : V \to W$ , since we are concentrating on $V \subseteq \mathbb{R}^n$, $W \subseteq \mathbb{R}^m$ we can represent T as $T(x) = Ax$.

   (a) An affine transformation is of the form :$S(x) = Bx + c$ and therefore every linear transformation is an affine transformation with $c = 0$.

   (b) An affine transformation with $c \neq 0$ cannot be represented as a linear transformation.

3. (a) The projection of vector v onto vector u:$P_{v,u} = \frac{\langle v,u \rangle}{\|u\|^2} \cdot u = \frac{-2+3+8}{\sqrt{1+1+4}} = -\frac{9}{\sqrt{6}}$

4.

5. Let $v, w \in \mathbb{R}^m, v, w \neq 0$. Let $\theta$ be the angle between $v$ and w.

$$0 = cos\theta \Leftrightarrow \frac{\langle u, v\rangle}{\|u\| \cdot \|v\|} = 0 \Leftrightarrow \langle u, v\rangle = 0$$

6. Let U be an orthogonal matrix and let A be an invertible matrix.$U, A \in \mathbb{R}^{nxn}$

$$E_A = \{y | y = Ax + v, \|x\| = 1, x \in \mathbb{R}^n\}, E_{AU} = \{y | y = AUx + v, \|x\| = 1, x \in \mathbb{R}^n\}$$

$\underline{E_{AU} \subseteq E_A :}$
$Let\, y \in E_{AU}$, $y = AUx+v, \|x\| = 1 \Rightarrow \langle x, x\rangle \Rightarrow \langle U^T Ux, x\rangle \Rightarrow \langle Ux, Ux\rangle \Rightarrow \|Ux\| = 1$
Now the affine transformation defining $E_A$on the vector$Ux$ gives$AUx+v = y \Rightarrow y \in E_A$
$\underline{E_A \subseteq E_{AU} :}$
Let $y \in E_A$ $y = Ax+v$, $\|x\| = 1$,the equation $Uz = x$ has a unique solution since U is invertible, therefore:

$$y = AUz + v, \|x\| = \|Uz\| = \|z\| = 1 \Longrightarrow y \in E_{AU}$$

7. Denote $L_{w,b} = \left\{x | w^T x = b\right\}$.
The projection of vector x in the hyperplane onto the vector w is:

$$P_{x,w} = \frac{x^T w}{\|w\|^2} w = \frac{b}{\|w\|^2} w$$

and the distance between the origin and the hyperplane is the norm of this projection :

$$Distance = \|P_{x,w}\| = \frac{b}{\|w\|^2} \|w\| = \frac{b}{\|w\|}$$

8.

9. Denote $A = \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$ and let $A = U\Sigma V^T$ be the SVD of A, denote the columns of U and V as $u_i$ and $v_i$ respectivly.

$$AA^T = U\Sigma V^T V \Sigma U^T = U\Sigma^2 U^T$$

$$AA^T = \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix} \cdot \begin{bmatrix} 5 & -1 \\ 5 & 7 \end{bmatrix} = \begin{bmatrix} 50 & 30 \\ 30 & 50 \end{bmatrix}$$

Let's find $AA^T$ eigenvalues by solving

$$det\left(AA^T - \lambda I\right) = 0$$

$det\left(\begin{bmatrix}50-\lambda & 30\\ 30 & 50-\lambda\end{bmatrix}\right) = (50-\lambda)^2 - 30^2 = 0 \Rightarrow 50^2 - 2\cdot 50\lambda + \lambda^2 - 30^2 \Rightarrow$
$\lambda^2 - 100\lambda + 1600$

$$\lambda_1 = 80\,,\ \lambda_2 = 20 \Rightarrow \sigma_1 = \sqrt{80}, \sigma_2 = \sqrt{20}$$

$$\begin{bmatrix}50-80 & 30\\ 30 & 50-80\end{bmatrix}\begin{bmatrix}u_{11}\\ u_{12}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \Rightarrow -30u_{11}+30u_{22} = 0 \Rightarrow u_{11} = u_{22}, u_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\ 1\end{bmatrix}$$

$$\begin{bmatrix}50-20 & 30\\ 30 & 50-20\end{bmatrix}\begin{bmatrix}u_{21}\\ u_{22}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \Rightarrow 30u_{11}+30u_{22} = 0 \Rightarrow u_{11} = -u_{22}, u_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\ -1\end{bmatrix}$$

$$U = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$$

$$A = U\Sigma V^T \Rightarrow U^T A = \Sigma V^T$$

therefore the i'th row of $\Sigma V^T$ is $\sigma_i v_i$ and is equal the the i'th row of $U^T A$ and so

$$U^T A = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\cdot\begin{bmatrix}5 & 5\\ -1 & 7\end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}4 & 12\\ 6 & -2\end{bmatrix}$$

$$\sigma_1 v_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}4\\ 12\end{bmatrix} \Rightarrow v_1 = \frac{4}{\sqrt{160}}\begin{bmatrix}1\\ 3\end{bmatrix} \Rightarrow v_1 = \frac{1}{\sqrt{10}}\begin{bmatrix}1\\ 3\end{bmatrix}$$

$$\sigma_2 v_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}6\\ -2\end{bmatrix} \Rightarrow v_2 = \frac{2}{\sqrt{40}}\begin{bmatrix}3\\ -1\end{bmatrix} \Rightarrow v_2 = \frac{1}{\sqrt{10}}\begin{bmatrix}3\\ -1\end{bmatrix}$$

$$A = U\Sigma V^T, U = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}, \Sigma = \begin{bmatrix}\sqrt{80} & 0\\ 0 & \sqrt{20}\end{bmatrix}, V = \frac{1}{\sqrt{10}}\begin{bmatrix}1 & 3\\ 3 & -1\end{bmatrix}$$

# Image Compression

10. The requested graphs:

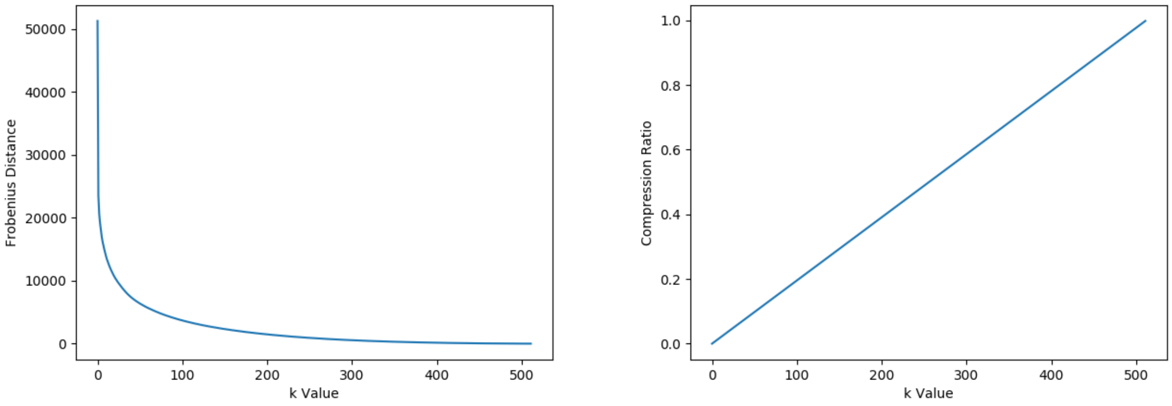


Figure 1: Frobenius Distance and Compression Ratio as a function of k

11. The requested compressed images:

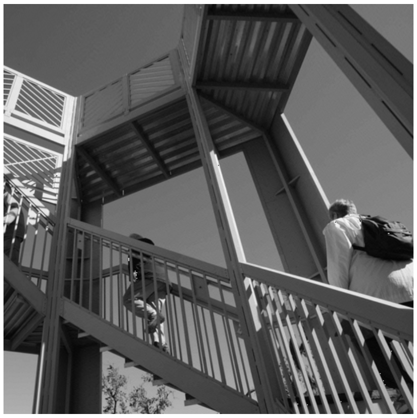


Figure 2: k value - 255 - Distance - 874.705229146026 - Compression ratio - 0.498046875

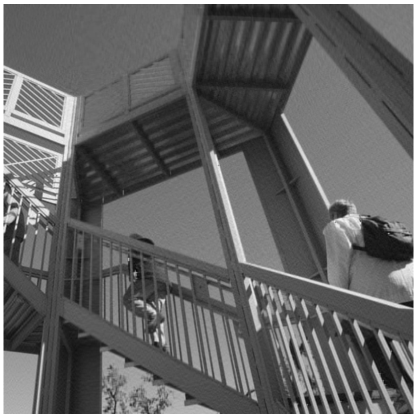


Figure 3: k value- 127 - Distance - 2851.5390926867926 - Compression Ratio - 0.248046875

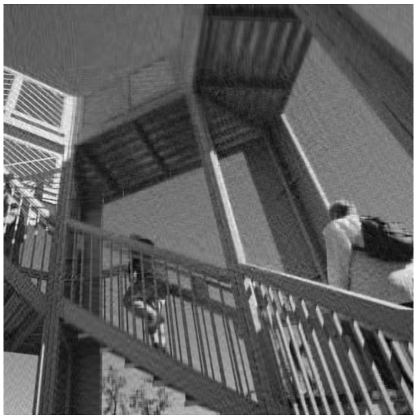


Figure 4: k value- 65 - Distance - 5329.527209019887 - Compression Ratio - 0.126953125

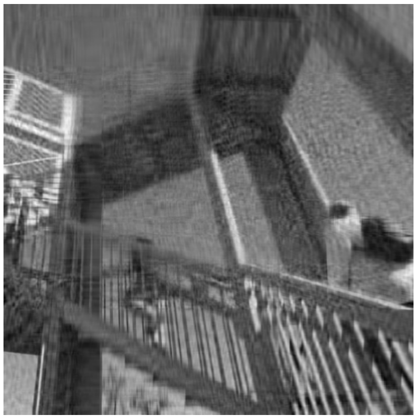


Figure 5: k value- 31 - Distance - 8545.535813063298- Compression Ratio -

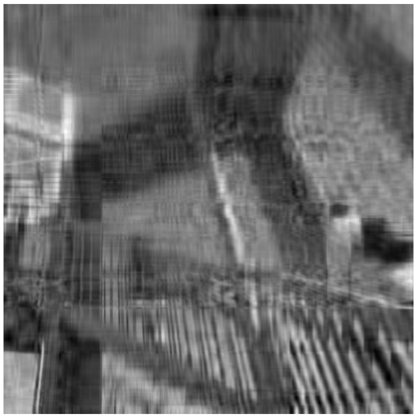


Figure 6: k value- 65 -
Distance - 5329.527209019887 -
Compression Ratio - 0.126953125

12. From the compression ratio graph we see that the compression ratio is linear in the number k and this is expected since each singular value we remove takes up the same space.
    From the frobenius distance graph we learn that the distance decreases very quickly when increasing the k value near 0 but then at some point of the graph the distance almost doesn't change as we increase k values.
    This means that the first singular values hold most of the image data and the higher ones hold almost no data.

# Linear Regression

## Theory: Normal Equations

13.

14. $\underline{Im(X^T) \subseteq Ker(X)^{\perp}:}$
    Let $v \in Im(X^T)$ therefore $\exists w$ s.t. $X^T w = v$. Now let $u \in Ker(X)$

    $$\langle u, v \rangle = \left\langle u, X^T w \right\rangle = \langle Xu, w \rangle = \langle 0, w \rangle = 0$$

    $\underline{Ker(X)^{\perp} \subseteq Im(X^T):}$
    Let $v \in Ker(X)^{\perp}$. Suppose, by contradiction, that $v \notin Im(X^T)$ then there exists a $v' \in Im\left(X^T\right)^{\perp}$ such that $\langle v, v' \rangle \neq 0$
    Since $X^T X v' \in Im(X^T)$ we get that $v' \in Ker(X)$. Therefore

    $$\langle Xv', Xv' \rangle = \left\langle v', X^T X v' \right\rangle = 0 \Rightarrow Xv' = 0 \Rightarrow v \notin (KerX)^{\perp}$$

In contradiction.

15. $X^T w = y$ , $X^T$ is not invertible
The system has $\infty$solutions iff $\mathrm{Ker}(X^T)$ is non trivial, i.e. it contains more than the 0 vector.
Denote one solution of the system as $w_1$. All other solutions can be represented as $w_1 + w_2$ where $w_2 \in Ker(X^T)$

$$X^T (w_1 + w_2) = y$$

Let $v \in Ker(X)$

$$v^T X^T (w_1 + w_2) = v^T y \Rightarrow (Xv)^T (w_1 + w_2) = v^T y$$

$$\Downarrow$$

$$v^T y = 0 \Rightarrow y \in Ker(X)^\perp$$

16. Consider the Normal linear system

$$XX^T w = Xy$$

If $XX^T$ is invertible than

$$w = Xy \left(XX^T\right)^{-1}$$

If $XX^T$ is uninvertible than according to article 15 the system

$$XX^T w = Xy$$

has $\infty$ solutions iff $Xy \perp Ker\left(XX^T\right)$ iff $Xy \perp Ker\left(X^T\right)$ since $Ker\left(X^T\right) = Ker\left(XX^T\right)$.
Let $u \in Ker\left(X^T\right)$

$$\langle Xy, u\rangle = \left\langle y, X^T u\right\rangle = \langle y, 0\rangle = 0$$

$$\Downarrow$$

$$Xy \perp Ker\left(X^T\right)$$

# Price Prediction

17. (a) I filtered any rows that have **ANY** null values.

(b) I defined all features that must be positive and all features that must be non-negative and filtered any rows that break any of these constraints.

| Feature | constraint | Reasoning |
|---|---|---|
| bedrooms | non-negative | Houses can have 0 bedrooms |
| bathrooms | non-negative | Houses can have 0 bathrooms |
| waterfront | non-negative | |
| view | non-negative | |
| sqft_above | non-negative | |
| sqft_basement | non-negative | |
| yr_built | non-negative | |
| yr_renovated | non-negative | |
| zipcode | non-negative | |
| sqft_living15 | non-negative | |
| sqft_lot15 | non-negative | |
| price | positive | |
| sqft_living | positive | |
| sqft_lot | positive | |
| floors | positive | |
| condition | positive | |
| grade | positive | |

(c) I removed the id column as it cannot have any possible influence on the house price.

(d) I added the interceptor feature.

(e) I removed the langtitude and longtitude features since I think the zipcode feature gives some information regarding the house location's influence on the price.
This is, however one area where I would invest more time if I had any to get better performance. I would unite these two features into a quantized feature of location and assume there is some linear relation between these locations and the house price.

(f) Regarding the date-time feature, I converted it to a number simply by taking the string representing the date (throwing the time since it is 0 for all samples) and converting it to a number. i.e '20140512T000000' is converted into 20140512.
The results with this date feature and without the date feature altogether do not seems to be distinctivly different.
Here, there is also extra work that can be done using the date feature, I would perhaps try to remove the day from the date and keep only the year and month, then trasform this feature to a one hot encoding feature.

18. I converted only the zip code feature to one hot encoded features, I did that using pandas get_dummies function.

19.

20.

21.

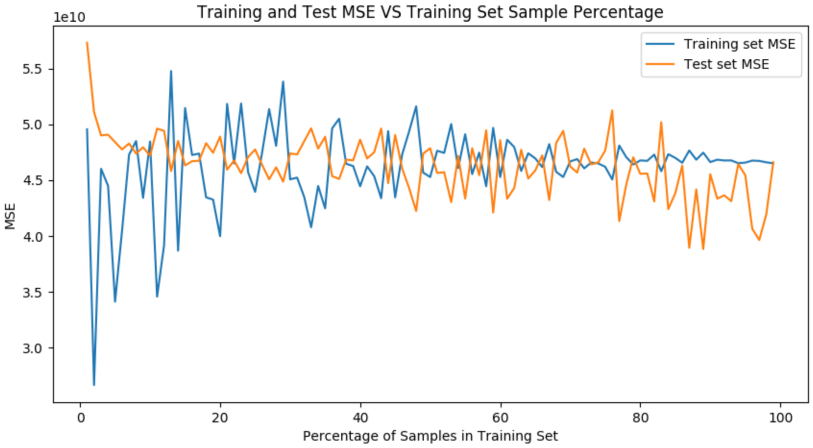


Figure 7: Training and Test MSE VS Training Samples Percentage - with date feature

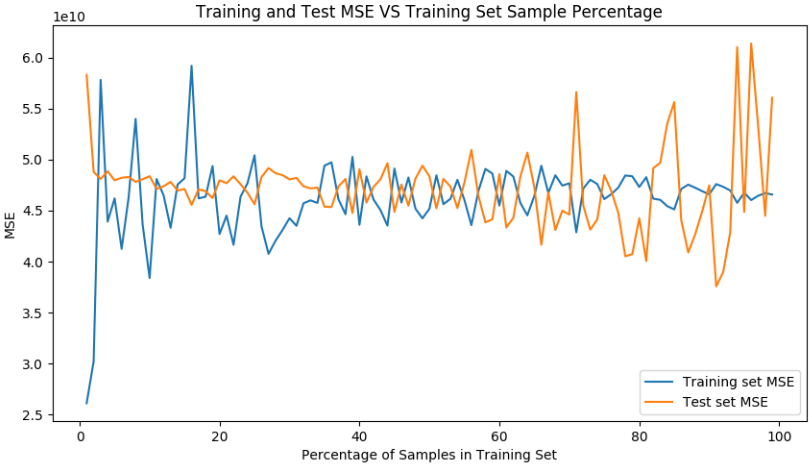


Figure 8: Training and Test MSE VS Training Samples Percentage - with date feature

# Concentration Inequalities

22. Chebyshev's inequality: The sample complexity is bounded above by $m(\epsilon, \delta) \leq \lceil \frac{1}{4\epsilon^2} \cdot \frac{1}{\delta} \rceil$

Hoeffding's inequality: The sample complexity is bounded above by $m(\epsilon, \delta) \leq \lceil \frac{1}{2\epsilon^2} \cdot log\left(\frac{2}{\delta}\right) \rceil$

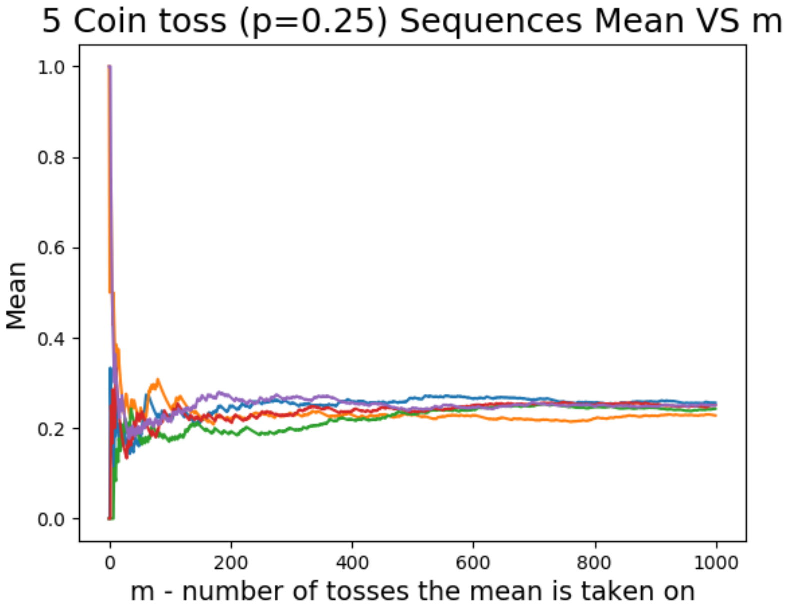
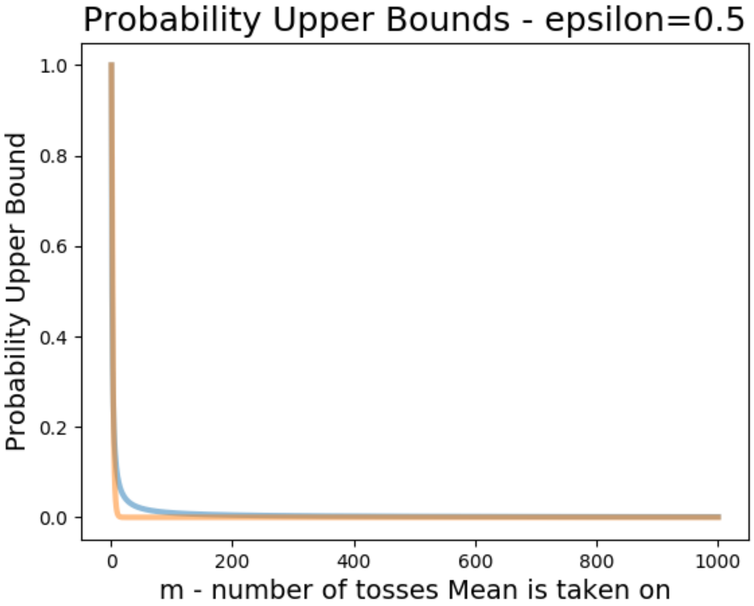
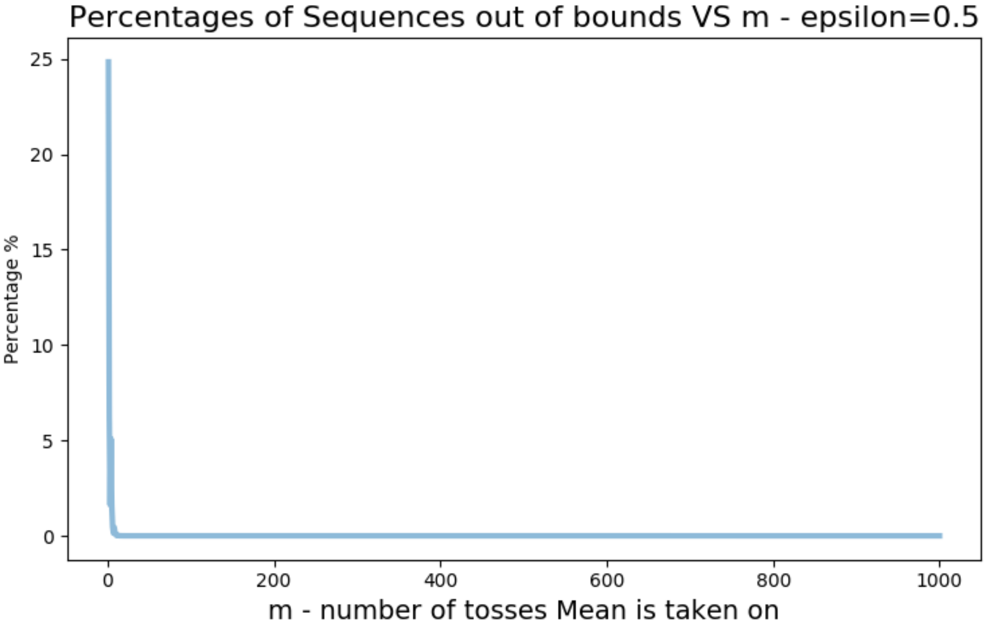


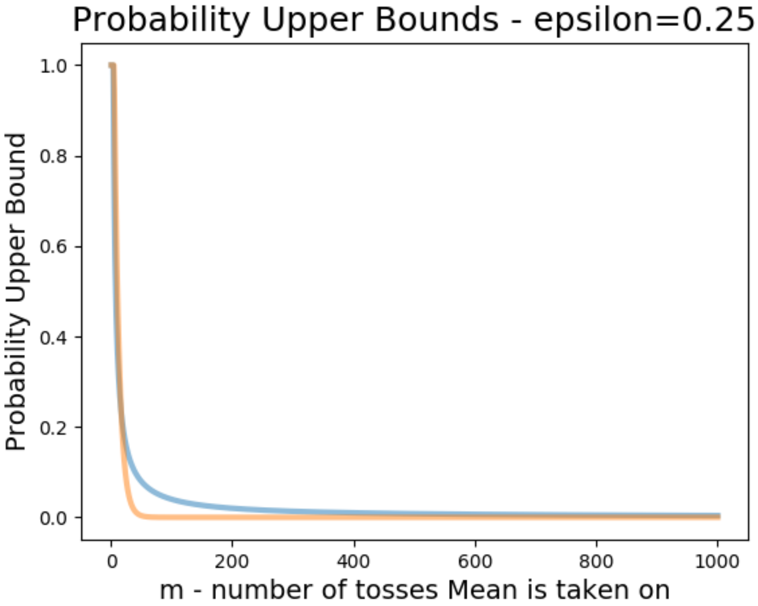
Figure 9: 5 Coin toss Sequences and their mean taken on first m tosses

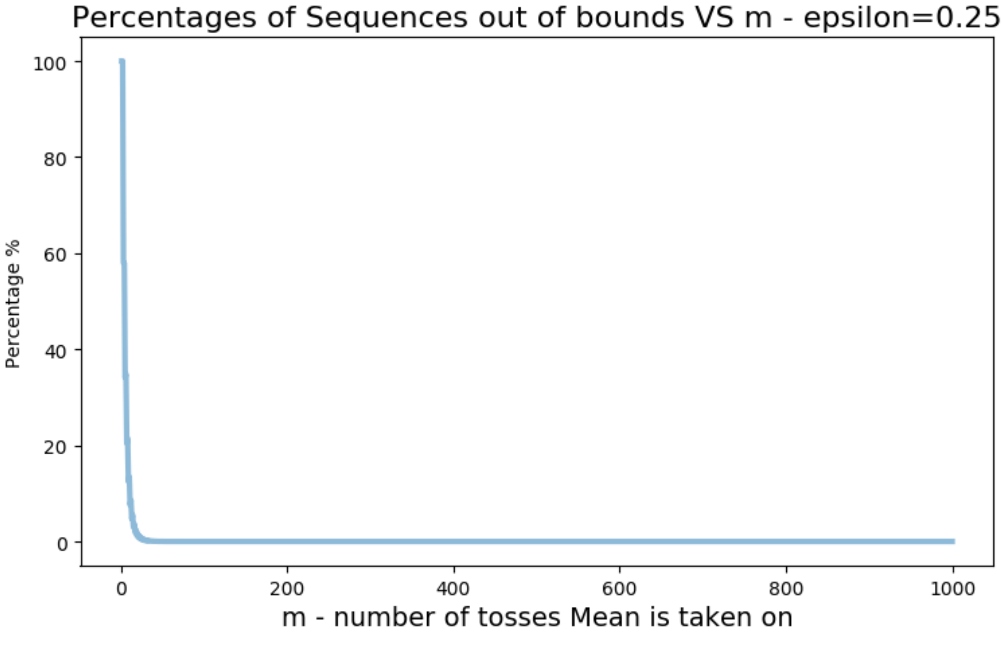
23. (a) I expect to see that as we increase m we get closer and closer to the true bias of the coin p=0.25 as is indeed the case for all 5 sequences.

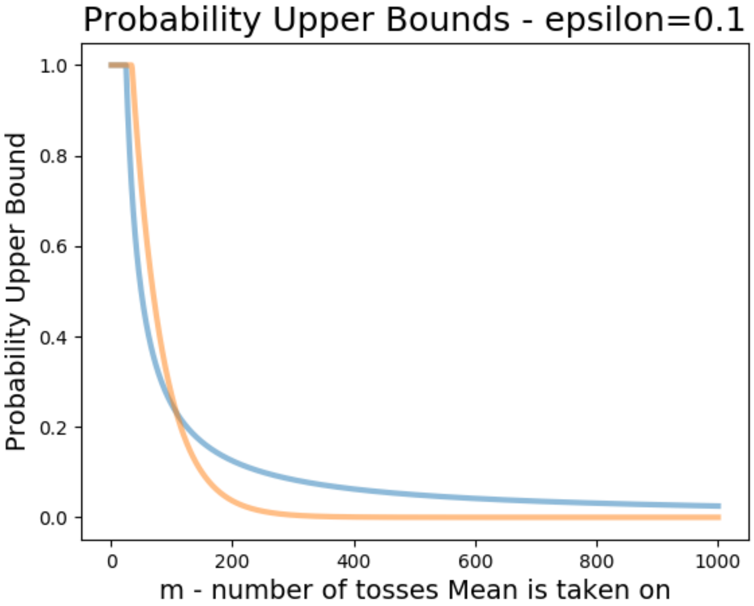
    (b) Following are the figures of the probability upper bounds

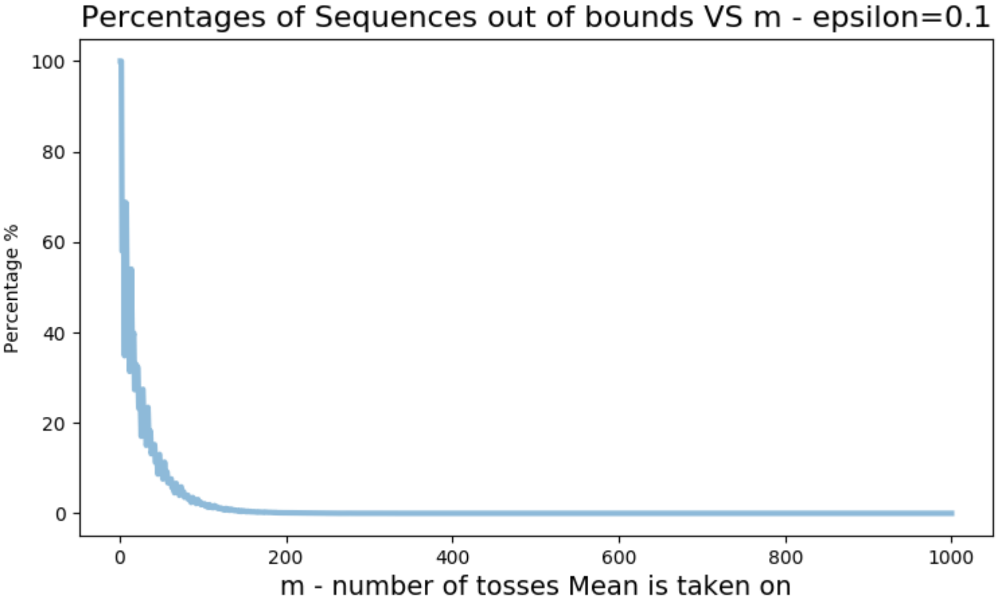

Probability Upper Bounds - epsilon=0.5
Probability Upper Bound
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
m - number of tosses Mean is taken on

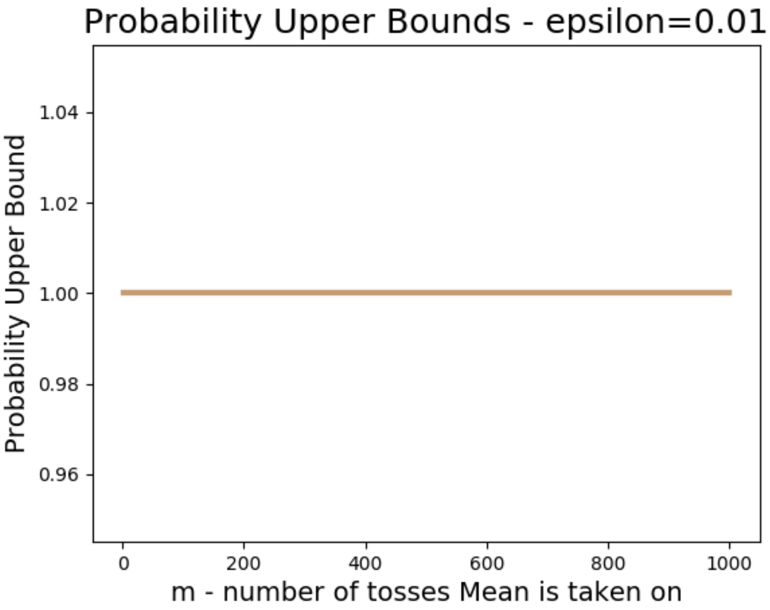

Percentages of Sequences out of bounds VS m - epsilon=0.5
Percentage %
25
20
15
10
5
0
0
200
400
600
800
1000
m - number of tosses Mean is taken on

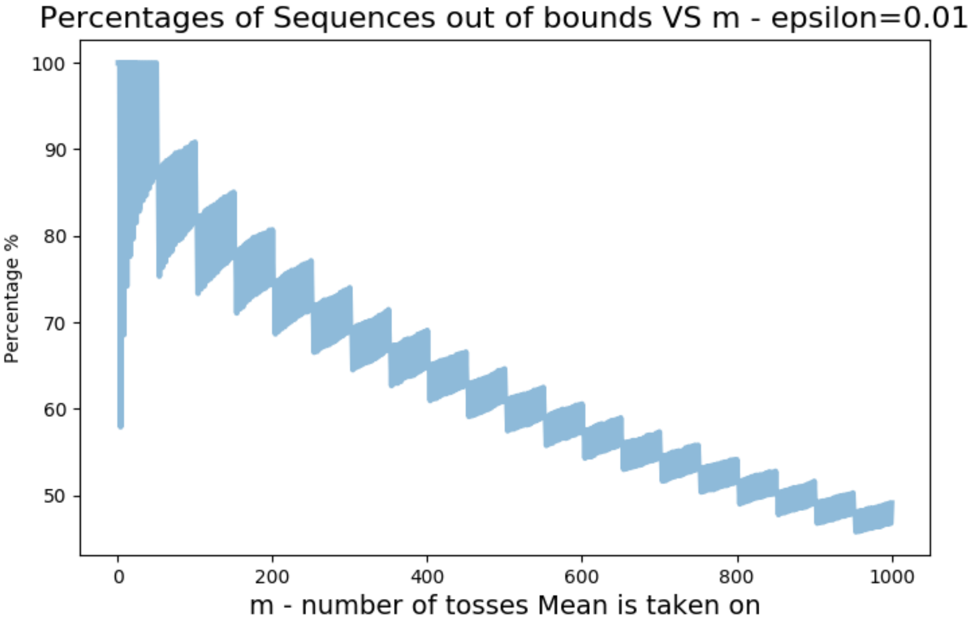

Probability Upper Bounds - epsilon=0.25
Probability Upper Bound
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
m - number of tosses Mean is taken on

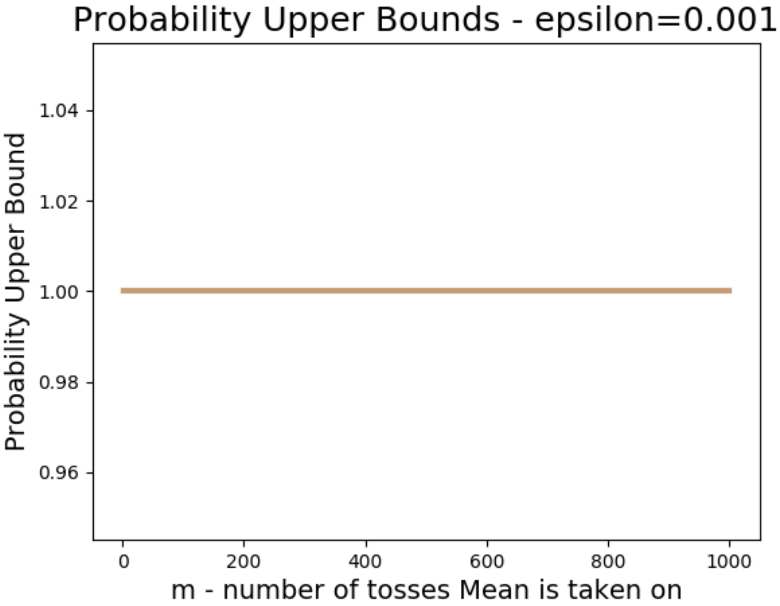

Percentages of Sequences out of bounds VS m - epsilon=0.25
Percentage %
100
80
60
40
20
0
0
200
400
600
800
1000
m - number of tosses Mean is taken on

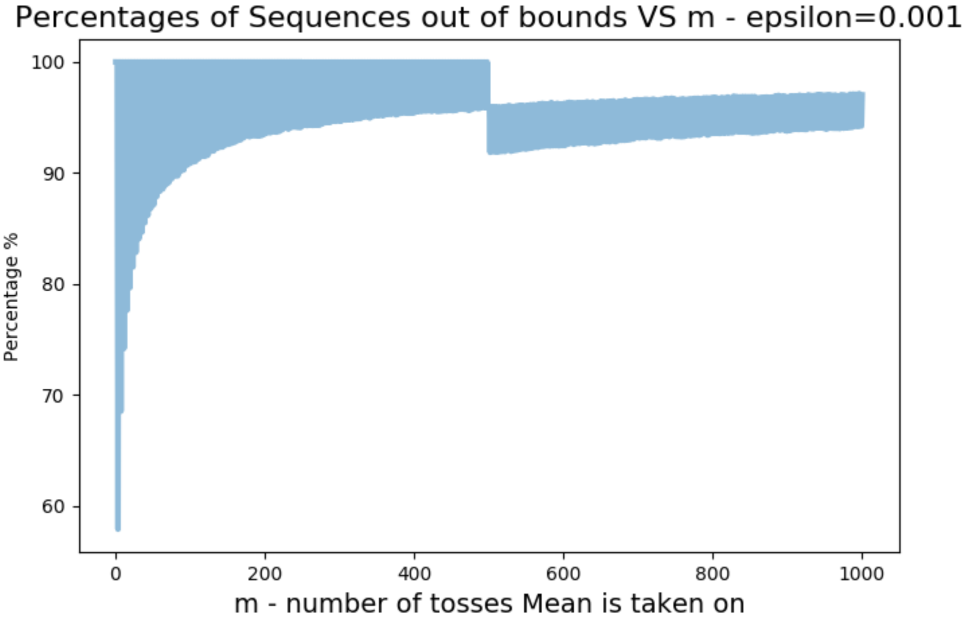

Probability Upper Bounds - epsilon=0.1
Probability Upper Bound
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
m - number of tosses Mean is taken on


Percentages of Sequences out of bounds VS m - epsilon=0.1
Percentage %
100
80
60
40
20
0
0
200
400
600
800
1000
m - number of tosses Mean is taken on


Probability Upper Bounds - epsilon=0.01
Probability Upper Bound
1.04
1.02
1.00
0.98
0.96
0
200
400
600
800
1000
m - number of tosses Mean is taken on


Percentages of Sequences out of bounds VS m - epsilon=0.01
Percentage %
100
90
80
70
60
50
0
200
400
600
800
1000
m - number of tosses Mean is taken on





(c) I expect that the percentage of sequences that satisfy having a distance of mean to expectation greater than $\epsilon$ will decrease as m in-

creases. This is because at a fixed $\epsilon$ the probability of a sequence having a distance of mean to expectation smaller than $\epsilon$ gets closer to 1 as m increases.
However, as $\epsilon$ decreases m will have to be bigger in order to decrease the probability that a sequence has a mean with a distance to the expectation greater than $\epsilon$.
Therefore, as $\epsilon$ decreases the percentage of sequences out of bounds will decrease only for greater values of m.