Introduction to Machine Learning (67577)

Exercise 1 - Mathematical Background

Dor Roter
208772251

March 27, 2021

1 Theoretical

1.1 Linear Algebra

1.1.1 Recap

1. $p = \frac{\langle v|w\rangle}{||w||^2} \cdot w = \frac{-2+3+8}{\sqrt{6}^2} \cdot \begin{pmatrix} 0 \\ -1 \\ 1 \\ 2 \end{pmatrix} = \frac{3}{2} w = \begin{pmatrix} 0 & -\frac{3}{2} & \frac{3}{2} & 3 \end{pmatrix}^T$

2. $p = \frac{\langle v|w\rangle}{||w||^2} \cdot w = \frac{1+3-4}{||w||^2} = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$

3. <u>“⇒”:</u>

$$\begin{aligned} &\text{let } 0 \neq v, w \in \mathbb{R}^m \text{so that } v \perp w \; (\theta = \pm 90^\circ) \\ &\Rightarrow \langle v|w\rangle = ||u|| \cdot ||w|| \cdot cos\left(\theta\right) = ||u|| \cdot ||w|| \cdot 0 = 0 \\ &\langle v|w\rangle := u^T w = 0 \quad \square \end{aligned}$$

<u>“⇐”:</u>

$$\begin{aligned} &v^T w =: \langle v|w\rangle = 0 \\ &\langle v|w\rangle = ||u|| \cdot ||w|| \cdot cos\left(\theta\right) = 0 \\ &\text{v, w are non-zero by def } \Rightarrow cos(\theta) = 0 \\ \Rightarrow &\theta = \pm 90^\circ \, \square \end{aligned}$$

4. let A be the corresponding matrix for $T : V \to W$ (a linear transforamtion), so that A is an orthogonal matrix, of size $n \times n$.

$$\begin{aligned} ||Ax|| &= \sqrt{\langle Ax|Ax\rangle} \underset{def}{=} \sqrt{\left(Ax\right)^T \cdot Ax} = \sqrt{x^T A^T Ax} = \sqrt{x^T x} = \\ &= \sqrt{\langle x|x\rangle} = ||x|| \qquad \square \end{aligned}$$

1.1.2 Singular Value Decomposition

5. let A be an inversable matrix ($\Rightarrow A \in \mathbb{R}^{n\times n}$) .
   $A = U\Sigma V^T$ (A's SVD decomposition)
   let $B = V\Sigma^{-1}U^T \Rightarrow A \cdot B = B \cdot A = I_n$
   U, V are orthogonal and therefore $\begin{matrix} UU^T = U^TU = I_n \\ VV^T = V^TV = I_n \end{matrix}$
   $\Rightarrow A \cdot B = U\Sigma V^T \cdot V\Sigma^{-1}U^T = U\Sigma I\Sigma^{-1}U^T = UIU^T = I_n$
   since Σ is diagonal and non-zero (else $A = 0$ which is not inversable), Σ^{-1}can easiliy be calcualted as well by inversing the singular value on the diagonal as:

   $$\begin{aligned} \Sigma &= diag\left(\sigma_1, ..., \sigma_n\right) \\ \Rightarrow\Sigma^{-1} &= diag\left(\sigma_1^{-1}, ..., \sigma_n^{-1}\right) \end{aligned}$$

   knowing the SVD of a matrix therefore enables us to find its inverse or even determine if the matrix is inversable without the need for complicated computations,
   simply by glancing at the SVD decomposition.

6. first we'll find A^TA's eigenvalues:

   $$A^TA = \begin{bmatrix} 5 & -1 \\ 5 & 7 \end{bmatrix} \cdot \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix} = \begin{bmatrix} 26 & 18 \\ 18 & 74 \end{bmatrix} \quad \text{(symetric $\Rightarrow$the P matrix of the EVD can be orthogonal)}$$
   $$det(A^TA - \lambda I_n) = det\left(\begin{bmatrix} 26-\lambda & 18 \\ 18 & 74-\lambda \end{bmatrix}\right) = 0$$
   $$\Rightarrow\lambda^2 - 100\lambda + 1600 = 0 \Rightarrow (\lambda - 80)(\lambda - 20) = 0$$
   $$\Rightarrow\lambda_1 = 80, \lambda_2 = 20$$

   no we can compute the eigenvectors corresponding to those:

   $$A^TA - 80I = \begin{bmatrix} -54 & 18 \\ 18 & -6 \end{bmatrix} \Rightarrow \left[\begin{array}{cc|c} -54 & 18 & 0 \\ 18 & -6 & 0 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & -\frac{1}{3} & 0 \\ 0 & 0 & 0 \end{array}\right]$$
   $$\Rightarrow x_1 = 1, x_2 = 3$$
   $$\Rightarrow v_1 = \frac{1}{\sqrt{10}}\begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

   $$A^TA - 20I = \begin{bmatrix} 6 & 18 \\ 18 & 54 \end{bmatrix} \Rightarrow \left[\begin{array}{cc|c} 6 & 18 & 0 \\ 18 & 54 & 0 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & 3 & 0 \\ 0 & 0 & 0 \end{array}\right]$$
   $$\Rightarrow x_1 = -3, x_2 = 1$$
   $$\Rightarrow v_2 = \frac{1}{\sqrt{10}}\begin{pmatrix} -3 \\ 1 \end{pmatrix}$$

   Therefore the EVD decomposition of A^TA is as such:

$$P = \frac{1}{\sqrt{10}}\begin{pmatrix}1 & -3\\ 3 & 1\end{pmatrix}, \quad D = \begin{pmatrix}80 & \\ & 20\end{pmatrix}, \quad P^{-1} = P^T = \frac{1}{\sqrt{10}}\begin{pmatrix}1 & 3\\ -3 & 1\end{pmatrix}$$

now, by applying some basic matrix operations using the SVD decomposition of A to define A^TA:

$$A^TA = V\Sigma\Sigma^TV^T \underset{\Sigma \text{ is diagonal}}{=} V\Sigma^2V^T = PDP^T$$

$$\Rightarrow\Sigma = \sqrt{D} = \begin{pmatrix}4\sqrt{5} & \\ & 2\sqrt{5}\end{pmatrix}, \quad V^T = P^T = \frac{1}{\sqrt{10}}\begin{pmatrix}1 & 3\\ -3 & 1\end{pmatrix}$$

as $A = U\Sigma V^T \underset{\text{multiply by } V}{\Leftrightarrow} AV = U\Sigma \underset{\text{multiply by } \Sigma^{-1}}{\Leftrightarrow} U = AV\Sigma^{-1}$:

$$U = \begin{pmatrix}5 & 5\\ -1 & 7\end{pmatrix} \cdot \frac{1}{\sqrt{10}}\begin{pmatrix}1 & -3\\ 3 & 1\end{pmatrix} \cdot \frac{1}{\sqrt{5}}\begin{pmatrix}\frac{1}{4} & \\ & \frac{1}{2}\end{pmatrix} = \frac{1}{5\sqrt{2}}\begin{pmatrix}5 & 5\\ -1 & 7\end{pmatrix}\begin{pmatrix}1 & -3\\ 3 & 1\end{pmatrix}\begin{pmatrix}\frac{1}{4} & \\ & \frac{1}{2}\end{pmatrix} =$$

$$= \frac{1}{5\sqrt{2}}\begin{pmatrix}20 & -10\\ 20 & 10\end{pmatrix}\begin{pmatrix}\frac{1}{4} & \\ & \frac{1}{2}\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$$

Therefore the SVD decomposition of A is $A = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix} \cdot \begin{pmatrix}4\sqrt{5} & \\ & 2\sqrt{5}\end{pmatrix} \cdot \frac{1}{\sqrt{10}}\begin{pmatrix}1 & 3\\ -3 & 1\end{pmatrix}$ □

7. let $A \in \mathbb{R}^{m\times n}$,then $C_0 = A^TA \in \mathbb{R}^{n\times n}$ a symetric and therfore diagnoizable matrix.

$$\forall k \in \mathbb{N} \quad b_{k+1} = \frac{C_0b_k}{||C_0b_k||}, \qquad b_0 = \sum_{i=1}^{n} a_iv_i = \begin{pmatrix}| & & |\\ v_1 & \cdots & v_n\\ | & & |\end{pmatrix}\begin{pmatrix}a_1\\ \vdots\\ a_n\end{pmatrix}$$

let us show using induction that for any $k \in \mathbb{N}$ it holds that $b_k = \frac{C_0^kb_0}{||C_0^kb_0||}$:

<u>**base:**</u> $b_1 := \frac{C_0b_0}{||C_0b_0||} = \frac{C_0^1b_0}{||C_0^1b_0||}$

<u>**step:**</u> let $k \in \mathbb{N}$ s.t $b_k = \frac{C_0^kb_0}{||C_0^kb_0||}$, then:

$$b_{k+1} := \frac{C_0b_k}{||C_0b_k||} = \frac{C_0 \cdot \frac{C_0^kb_0}{||C_0^kb_0||}}{||C_0 \cdot \frac{C_0^kb_0}{||C_0^kb_0||}||} \underset{\text{norm homogeneity}}{=} \frac{\frac{1}{||C_0^kb_0||} \cdot C_0^{k+1}b_0}{\frac{1}{||C_0^kb_0||} \cdot ||C_0^{k+1}b_0||} = \frac{C_0^{k+1}b_0}{||C_0^{k+1}b_0||}$$

let $v_1, \ldots, v_n \in \mathbb{R}^n$ be eigenvectors of C_0 corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$.

let $P = \begin{pmatrix}| & & |\\ v_1 & \cdots & v_n\\ | & & |\end{pmatrix}$, $D = \left(\begin{array}{ccc|c}\lambda_1 & & & \\ & \ddots & & 0\\ & & \lambda_n & \\ \hline & 0 & & 0\end{array}\right)$, therefore using the EVD decomposition of C_0 it holds that: $C_0 = PDP^{-1} \Rightarrow C_0^k = PDP^{k-1}$

from P's definition:

$$(i) P\cdot e_i = v_i \Leftrightarrow P^{-1}\cdot v_i = e_i$$
$$(ii) P\cdot \alpha e_i = \alpha v_i \quad (\alpha\in\mathbb{R})$$

Therefore:

$$\Rightarrow C_0^k b_0 \underset{b_0\text{def}}{=} \sum_{i=1}^{n} C_0^k a_i v_i \underset{C_0\text{def}}{=} \sum_{i=1}^{n} a_i\cdot PD^kP^{-1}v_i \underset{(i)}{=} \sum_{i=1}^{n} a_i\cdot PD^k\cdot e_i \underset{D\text{ def}}{=}$$
$$= \sum_{i=1}^{n} a_i\cdot P\lambda_i^k\cdot e_i \underset{(ii)}{=} \sum_{i=1}^{n} a_i\lambda_i^k v_i = a_1\lambda_1\cdot\left(v_1+\sum_{i=2}^{n}\frac{a_i\lambda_i}{a_1\lambda_1}v_i\right)$$

$$\Rightarrow ||C_0^k b_0|| = \sqrt{||C_0^k b_0||^2} \underset{\substack{pythagorean\\ theorem}}{=} \sqrt{\sum_{i=1}^{n}||a_i\lambda_i^k v_i||^2} \underset{\text{norm homogeneity}}{=}$$
$$= \sqrt{\sum_{i=1}^{n}\left|a_i\lambda_i^k\right|^2||v_i||^2} = |a_1\lambda_1|\cdot\left(||v_1||+\sum_{i=2}^{n}\left(\frac{a_i\lambda_i}{a_1\lambda_1}\right)^2||v_i||^2\right)$$

tying it all together we get:

$$\lim_{k\to\infty} b_k = \lim_{k\to\infty}\frac{C_0^k b_0}{||C_0^k b_0||} = \lim_{k\to\infty}\frac{a_1\lambda_1\cdot\left(v_1+\sum_{i=2}^{n}\frac{a_i\lambda_i}{a_1\lambda_1}v_i\right)}{|a_1\lambda_1|\cdot\left(||v_1||+\sum_{i=2}^{n}\left(\frac{a_i\lambda_i}{a_1\lambda_1}\right)^2||v_i||^2\right)}$$

since for each $i\in[n]$ $\lambda_1>\lambda_2\geq\cdots\geq\lambda_n$, it holds that for each $i>1$: $\frac{\lambda_i}{\lambda_1}<1\Rightarrow\lim_{k\to\infty}\left(\frac{\lambda_i}{\lambda_1}\right)^k=0$

$$\lim_{k\to\infty} b_k = \lim_{k\to\infty}\frac{a_1\lambda_1\cdot\left(v_1+\sum_{i=2}^{n}\frac{a_i\lambda_i}{a_1\lambda_1}v_i\right)}{|a_1\lambda_1|\cdot\left(||v_1||+\sum_{i=2}^{n}\left(\frac{a_i\lambda_i}{a_1\lambda_1}\right)^2||v_i||^2\right)} = \frac{a_1\lambda_1\cdot(v_1+0)}{|a_1\lambda_1|\cdot(||v_1||+0)} \underset{||v_i||=1}{=} \pm v_i \qquad \square$$

1.2 Multivariate Calculus

8. let $f(\sigma) = U \cdot diag(\sigma) \cdot U^T x$ where $U \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$
first we will single out a single value out of f's image:

$$f(\sigma) = U \cdot diag(\sigma) \cdot U^T x = U\sigma^T U^T x$$

$$\Rightarrow f_i(\sigma) = u_i \cdot \left(diag(\sigma) U^T x\right) = u_i \cdot \langle \sigma | U^T x \rangle = u_i \cdot \begin{pmatrix} \sigma_1 \left(U^T x\right)_1 \\ \vdots \\ \sigma_n \left(U^T x\right)_n \end{pmatrix} = \sum_{j=1}^{n} u_i^j \sigma_j \cdot \left(U^T x\right)_j$$

now computing f_i's partial derivitives:

$$\frac{\partial f_i(\sigma)}{\partial \sigma_j} = \sum_{j=1}^{n} \frac{\partial}{\partial \sigma_j} \left(u_i^j \sigma_j \cdot \left(U^T x\right)_j \right) = u_i^j \sigma_j \left(U^T x\right)_j$$

$$\nabla f_i(\sigma)^T = \begin{pmatrix} u_i^1 \sigma_1 \left(U^T x\right)_1 & \cdots & u_i^n \sigma_n \left(U^T x\right)_n \end{pmatrix}$$

$$J_\sigma(f) = \begin{pmatrix} - & \nabla f_1(\sigma)^T & - \\ & \vdots & \\ - & \nabla f_n(\sigma)^T & - \end{pmatrix} = \begin{pmatrix} u_1^1 \sigma_1 \left(U^T x\right)_1 & \cdots & u_1^n \sigma_n \left(U^T x\right)_n \\ \vdots & & \vdots \\ u_n^1 \sigma_1 \left(U^T x\right)_1 & \cdots & u_n^n \sigma_n \left(U^T x\right)_n \end{pmatrix} = U \cdot diag\left(U^T x\right) \qquad \square$$

9. let $h(\sigma) = \frac{1}{2}||f(\sigma) - y||^2$, $g(x) = ||x||^2$ and $f(\sigma) = U \cdot diag(\sigma) \cdot U^T x$.
then we can rewrite h as: $h := \frac{1}{2} \cdot g \circ (f - y)$.
now, using the chain rule:

$$J_\sigma(h) = J_\sigma\left(\frac{1}{2} \cdot g \circ (f-y)\right) = \frac{1}{2} \cdot J_{(f-y)\sigma}(g) \cdot J_\sigma(f-y) \underset{\substack{\text{as seen} \\ \text{in recetaion}}}{=} \left[2\left(f(\sigma) - y\right]^T \cdot J_\sigma(f) =$$

$$= \left(U \cdot diag(\sigma) \cdot U^T x - y\right)^T U \cdot diag\left(U^T x\right) = \nabla MSE(f) \qquad \square$$

10. let $g(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{k} e^{z_k}}$

$$\frac{\partial}{\partial z_j} g(z)_i = \frac{\partial}{\partial z_j}\left(\frac{e^{z_i}}{\sum_{k=1}^{k} e^{z_k}}\right) = \frac{\frac{\partial e^{z_i}}{\partial z_j} \cdot \left(\sum_{k=1}^{k} e^{z_k}\right) - e^{z_i} \cdot \frac{\partial\left(\sum_{k=1}^{k} e^{z_k}\right)}{\partial z_j}}{\left(\sum_{k=1}^{k} e^{z_k}\right)^2}$$

$$= \begin{cases} \frac{e^{z_i} \cdot \left(\sum_{k=1}^{k} e^{z_k}\right) - e^{z_i} \cdot e^{z_j}}{\left(\sum_{k=1}^{k} e^{z_k}\right)^2} = \left(\frac{e^{z_i}}{\sum_{k=1}^{k} e^{z_k}}\right) \cdot \left(\frac{\left(\sum_{k=1}^{k} e^{z_k}\right) - e^{z_j}}{\left(\sum_{k=1}^{k} e^{z_k}\right)^2}\right) = g(z)_i \cdot (1 - g(z)_j) & i = j \\ \frac{-e^{z_i} \cdot e^{z_j}}{\left(\sum_{k=1}^{k} e^{z_k}\right)^2} = -\left(\frac{e^{z_i}}{\sum_{k=1}^{k} e^{z_k}}\right) \cdot \left(\frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}}\right) = -g(z)_i \cdot g(z)_j = g(z)_i \cdot (0 - g(z)_j) & i \neq j \end{cases}$$

therefore we can use Kronecker delta to represent the expression as:

$$J_z(g) = \begin{pmatrix} \nabla g(z)_1 \\ \vdots \\ \nabla g(z)_K \end{pmatrix} \Rightarrow J_z(g)_i^j = \frac{\partial g(z)_i}{\partial z_j} = g(z)_i \cdot \left(\delta_i^j - g(z)_j\right) \qquad \square$$

2 Practical

2.1 Multivariate Gaussians

11. based on the covariance and mean used to generate the dataset we can deduce the dataset is of uncorrelated variables with identical variance (there are zero values off diagonal, and roughly one for the values on the diagonal of the covariance matrix).

As expected the plot resembles a standard gaussian:

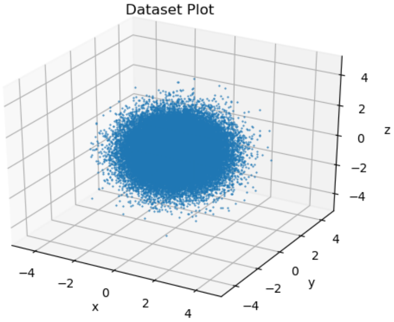
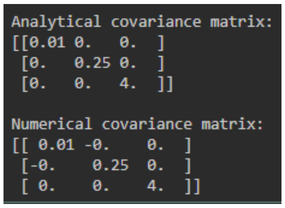


Figure 1: scatter of the generated gaussian distirbuted 3D dataset

12. Having a look at the covariance matrix, computed both analyticly and numerically:



```
Analytical covariance matrix:
[[0.01 0.   0.  ]
 [0.   0.25 0.  ]
 [0.   0.   4.  ]]

Numerical covariance matrix:
[[ 0.01 -0.    0.  ]
 [-0.    0.25  0.  ]
 [ 0.    0.    4.  ]]
```

We can notice that the data point of the X and Y axis compacted while the Z axes stretched. as the scaling matrix is diagonal, the X,Y and Z axis scaling is uncorrelated, since the off diagonal elements of the variance matrix are in fact roughly zero. We can also tell how closely the analytical and numerical calculations of the covariance matrix is thanks to the size of the dataset.

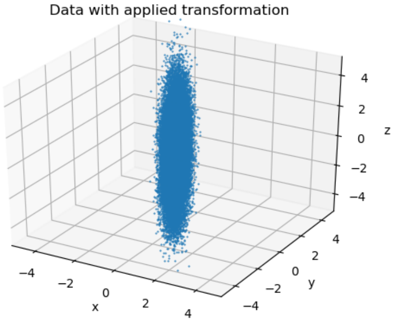


Figure 2: scatter of the scaled dataset

13. We note, as expected, that the orthogonal transformation in fact preformed just a rotation of the dataset without stretching it. Yet we do notice the covariance matrix indicate there is some correlation between the 3 axis.

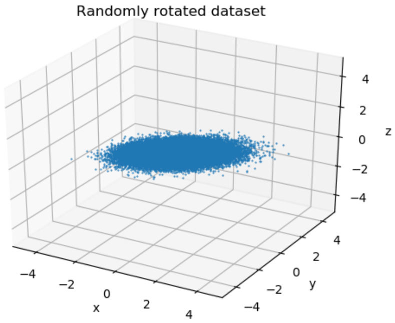
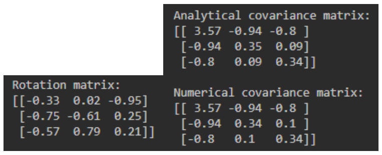


Figure 3: scatter of the rotated dataset



```
Analytical covariance matrix:
[[ 3.57 -0.94 -0.8 ]
 [-0.94  0.35  0.09]
 [-0.8   0.09  0.34]]

Rotation matrix:
[[-0.33  0.02 -0.95]
 [-0.75 -0.61  0.25]
 [-0.57  0.79  0.21]]

Numerical covariance matrix:
[[ 3.57 -0.94 -0.8 ]
 [-0.94  0.34  0.1 ]
 [-0.8   0.1   0.34]]
```

(a) covariance & rotation matrices

14. as expected, the marginal distribution of the data clearly resembles the normal distribution for both the X and Y axis, and as such seems to indeed be a gaussian, although it is slightly stretched across the X axis and shrunk across the Y axis, which is to be expected considering the scaling matrix we have applied to the data.

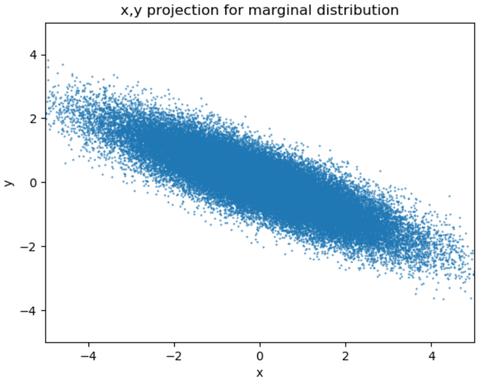
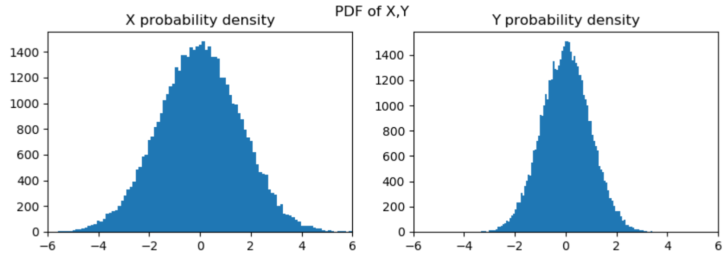


Figure 4: projection for marginal distribution of a gaussian



(a) histograms of the projection

15. as expected, the conditional distribution of the data clearly resembles the normal distribution for both the X and Y axis, and as such seems to indeed be a gaussian, although it is slightly stretched across the X axis and shrunk across the Y axis, which is to be expected considering the scaling matrix we have applied to the data.

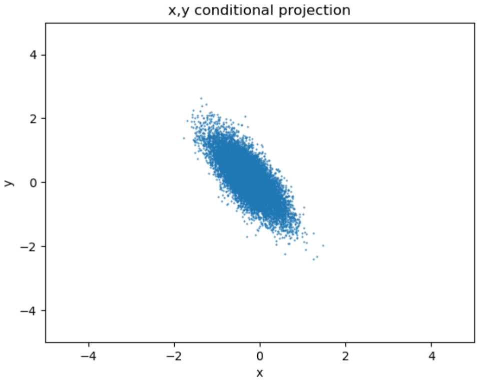
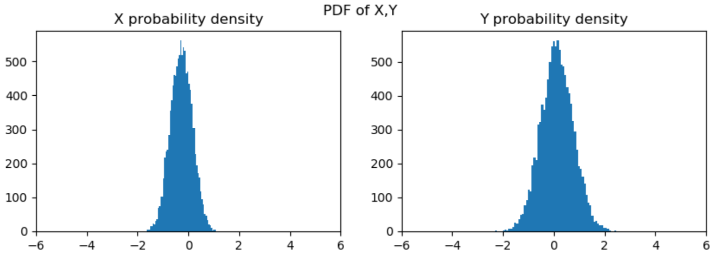


Figure 5: projection for conditional distribution of a gaussian



(a) histograms of the projection

2.2 Concentration Inequalities

16. (a). As stated by the law of large numbers, the larger the sample size (m), the closer the average $(\overline{X}_m)$ will be to actual mean of the coin - which seems to be roughly $\frac{1}{4}$. We can clearly see the value of each of the sequences converging to the same value, which corresponds, as expected, to the fact all sequences were created using the same "coin".

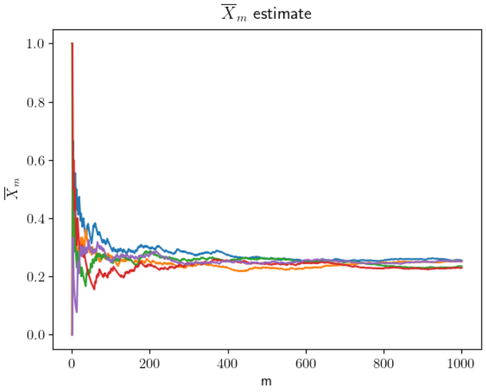


Figure 6: estimation visualization

(b). Since we know the dataset is of some p-biased coin: $X_i \sim Ber(p) \Rightarrow Var(X_i) = p \cdot (1-p) \leq \frac{1}{4}$. Therefore, we can now analytically calculate both upper bound - Chebyshev's and Hoeffding's, as shown in the recitation.

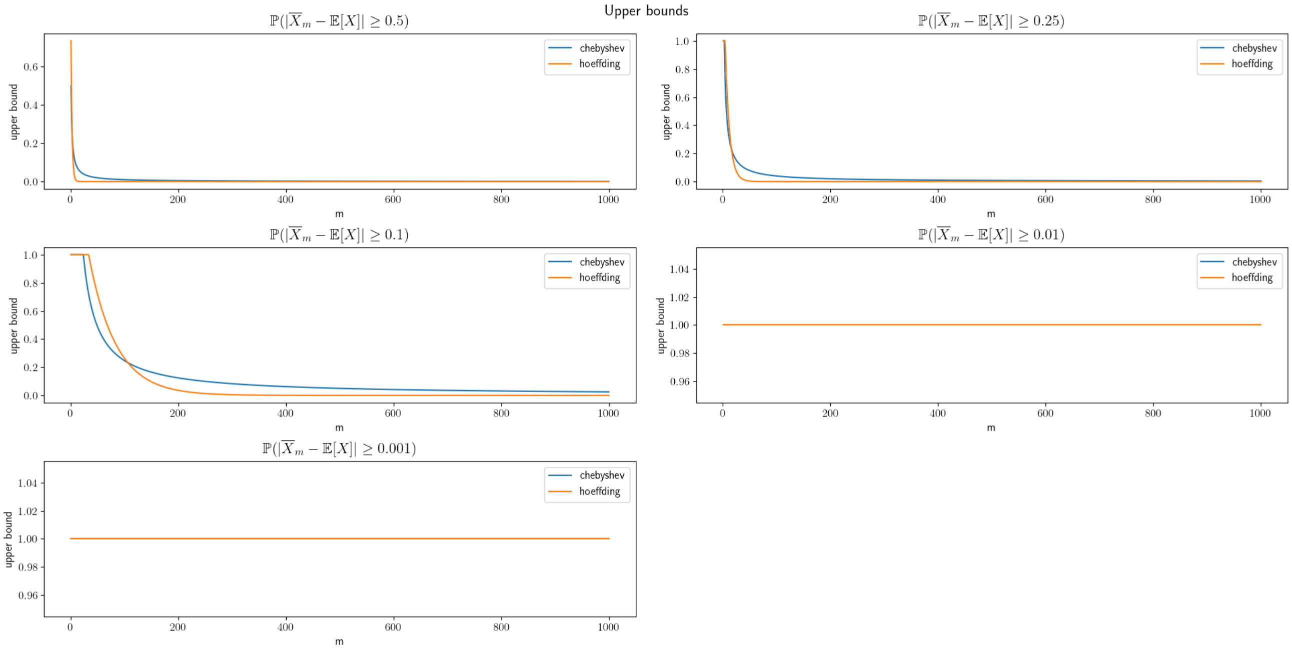


Figure 7: Upper bounds using different epsilons

(c). Now that we know as a matter of fact $\mathbb{E}[X] = 0.25$, I was expecting to see the plots converging to the value 0 as m grows larger, while also the larger the epsilon value is, the faster the plot will converge to 0 - meaning almost all the computed $\overline{X}_m$ sequences values (estimated means) are in the epsilon range of the actual mean. So we can see as a matter of fact:

$$\forall \epsilon > 0 \ \lim_{m \to \infty} \mathbb{P}(|\overline{X}_m - \mathbb{E}(X)| > \epsilon) = 0$$

Exactly as stated by the law of large numbers.

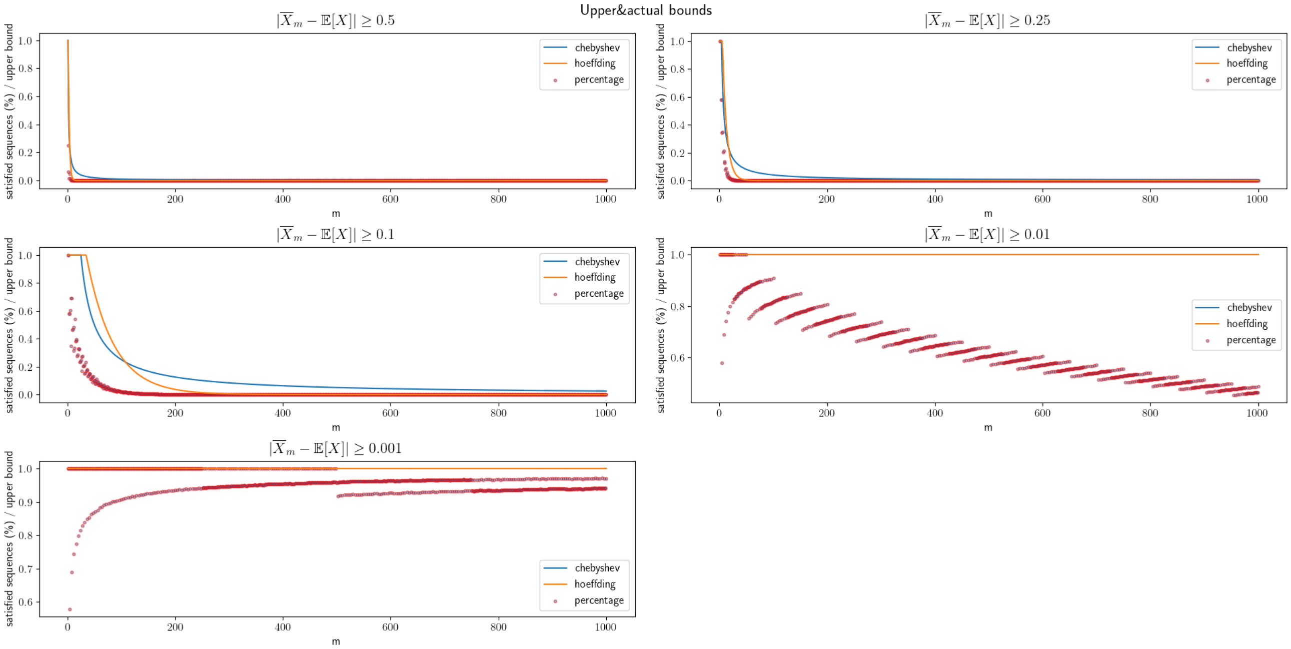


Figure 8: actual bounds using different epsilons