Introduction to Machine Learning (67577)
Exercise I – Mathematical Background
Hadar Sharvit – 208287599

PART I

Question 1 – Linear Algebra

1. $\vec{p} = \frac{\vec{w}\cdot\vec{v}}{|\vec{w}|^2}\vec{w} = \frac{9}{6}\vec{\omega} = (0 \quad -3/2 \quad 3/2 \quad 3)$
2. Since $v \cdot w = 1 + 0 + 4 - 4 = 0$ we have $\vec{p} = \vec{0} = (0 \quad 0 \quad 0 \quad 0)$
3. In the first direction - let $v, w \in \mathbb{R}^m$ be two non-zero vectors with angle $\theta = \pm\pi/2$ between them. This means that $\cos\theta = 0 = \frac{\vec{w}\cdot\vec{v}}{||w||\cdot||v||}$, i.e. $v \cdot w = v^T w = 0$

   In the second direction – if $v^T w = v \cdot w = ||v||||w||cos\theta = 0$, and since the norms > 0 it must be that $cos\theta = 0 \rightarrow \theta = \pm\pi/2$
4. $||Ax||^2 = (Ax)^T(Ax) = x^T A^T A x = x^T I x = x^T x = ||x||^2 \rightarrow ||Ax|| = ||x||$
5. Denoting $A = U\Sigma V^T$ where U, V are orthonormal matrices and Σ is a diagonal matrix with values ≥ 0 we can write
   $$A^{-1} = (U\Sigma V^T)^{-1} = (V^T)^{-1}\Sigma^{-1}U^{-1}$$
   Since V is orthogonal then $V^T = V^{-1} \rightarrow (V^T)^{-1} = V$ and we get
   $$A^{-1} = V\Sigma^{-1}U^{-1}$$
   This is useful because it only takes two calculation of transpose in order to find A^{-1}
6. $A = U\Sigma V^T = \begin{pmatrix} 5 & 5 \\ -1 & 7 \end{pmatrix}$
   $$A^T A = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^T U^T U \Sigma V^T = V\Sigma^T\Sigma V^T = V\Sigma^2 V^{-1}$$
   - $A^T A = \begin{pmatrix} 5 & -1 \\ 5 & 7 \end{pmatrix}\begin{pmatrix} 5 & 5 \\ -1 & 7 \end{pmatrix} = \begin{pmatrix} 26 & 18 \\ 18 & 74 \end{pmatrix}$
   - To find the eigenvalues we calculate $\det(A^T A - \lambda I) \overset{!}{=} 0$:
   $$\begin{vmatrix} 26-\lambda & 18 \\ 18 & 74-\lambda \end{vmatrix} = (26-\lambda)(74-\lambda) - 18^2 \overset{!}{=} 0 \rightarrow \lambda_{1,2} = 20{,}80$$
   Therefore, we can write
   $$\Sigma = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & \sqrt{20} \end{pmatrix} = \sqrt{10}\begin{pmatrix} 2\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{pmatrix}$$
   - Now for the eigenvectors:

   For $\lambda_1 = 20$ we have
   $$A^T A - 20I = \begin{pmatrix} 6 & 18 \\ 18 & 54 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = 0 \rightarrow \begin{matrix} a + 3b = 0 \\ 3a + 9b = 0 \end{matrix} \rightarrow a = -3b, b \in \mathbb{R}$$
   So, we can conclude $v_2 || \begin{pmatrix} -3 \\ 1 \end{pmatrix}$ and after normalization ($||v_2|| = \sqrt{10}$)
   $$v_2 = \frac{1}{\sqrt{10}}\begin{pmatrix} -3 \\ 1 \end{pmatrix}$$
   For $\lambda_2 = 80$ we have
   $$A^T A - 20I = \begin{pmatrix} -54 & 18 \\ 18 & -6 \end{pmatrix} \rightarrow \begin{pmatrix} -9 & 3 \\ 3 & -1 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = 0 \rightarrow \begin{matrix} -9a + 3b = 0 \\ 3a - b = 0 \end{matrix} \rightarrow a = \frac{b}{3}, b \in \mathbb{R}$$
   So, we can conclude $v_1 || \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ and after normalization:
   $$v_1 = \frac{1}{\sqrt{10}}\begin{pmatrix} 1 \\ 3 \end{pmatrix}$$
   This gives us

$$V = \frac{1}{\sqrt{10}}\begin{pmatrix}1 & -3\\ 3 & 1\end{pmatrix}$$

- Finally, since $AV = U\Sigma$ then $U = AV\Sigma^{-1}$:

$$U = \begin{pmatrix}5 & 5\\ -1 & 7\end{pmatrix}\frac{1}{\sqrt{10}}\begin{pmatrix}1 & -3\\ 3 & 1\end{pmatrix}\frac{1}{\sqrt{10}}\begin{pmatrix}2\sqrt{2} & 0\\ 0 & \sqrt{2}\end{pmatrix}^{-1} = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$$

Let's check everything:

$$U\Sigma V^T = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}\sqrt{10}\begin{pmatrix}2\sqrt{2} & 0\\ 0 & \sqrt{2}\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}}\\ \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}}\end{pmatrix}^T = \begin{pmatrix}5 & 5\\ -1 & 7\end{pmatrix} = A$$

7. $b_{k+1} = C_0 b_k / ||C_0 b_k||\ \forall k \in \mathbb{N}$, $C_0 = A^T A$ where $v_1, \dots v_n$ and $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n$ are the eigenvectors and eigenvalues of A respectively.
Iteratively expanding b_k:

$$b_k = \frac{C_0 b_{k-1}}{||C_o b_{k-1}||} = \frac{C_0^2 b_{k-2}}{||C_0^2 b_{k-2}||} = \dots = \frac{C_0^k b_0}{||C_0^k b_0||}$$

Using the hint for b_0:

$$\frac{C_0^k b_0}{||C_0^k b_0||} = \frac{C_0^k \sum_{i=1}^n a_i v_i}{||C_0^k \sum_{i=1}^n a_i v_i||}$$

Using C_0's *EVD* we have:

$$C_0 = A^T A = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T$$

Where $\Sigma^T\Sigma$ is a diagonal matrix in which $(\Sigma^T\Sigma)_{ii}$ are the eigenvalues of C_0. From here, as we've seen in class

$$C_0^k = V(\Sigma^T\Sigma)^k V^T$$

This can be substituted to b_k in summation form as followed

$$b_k = \frac{\sum_{i=1}^n a_i C_0^k v_i}{||\sum_{i=1}^n a_i C_0^k v_i||} = \frac{\sum_{i=1}^n a_i \lambda_i^k v_i}{||\sum_{i=1}^n a_i \lambda_i^k v_i||} = \frac{a_1\lambda_1^k\left(v_1 + \sum_{i=2}^n \frac{a_i}{a_1}\left(\frac{\lambda_i}{\lambda_1}\right)^k v_i\right)}{\left|\left|a_1\lambda_1^k\left(v_1 + \sum_{i=2}^n \frac{a_i}{a_1}\left(\frac{\lambda_i}{\lambda_1}\right)^k v_i\right)\right|\right|}$$

Since $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n$ then $\lim_{k\to\infty}\left(\frac{\lambda_i}{\lambda_1}\right)^k = 0$ so we can write

$$\lim_{k\to\infty} b_k = \frac{a_1\lambda_1^k v_1}{||a_1\lambda_1^k v_1||} \overset{(\dagger)}{=} \pm v_1$$

(†) this is because $|v_1| = 1$ so $\frac{v_1}{||v_1||} = v_1$ and the sign depends on the value of $a_1\lambda_1^k$

Question 2 – Multivariable Calculus

8. $f(\sigma) = U \cdot diag(\sigma)U^T x = \begin{pmatrix}u_{11} & \cdots & u_{1n}\\ \vdots & \ddots & \vdots\\ u_{n1} & \cdots & u_{nn}\end{pmatrix}\begin{pmatrix}\sigma_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_n\end{pmatrix}\begin{pmatrix}u_{11} & \cdots & u_{n1}\\ \vdots & \ddots & \vdots\\ u_{1n} & \cdots & u_{nn}\end{pmatrix}\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix}$
Which can be written as $f(\sigma) = \sum_{i=1}^n \sigma_i u_i u_i^T x$ where i represents the $i'th$ column.
From here we can calculate

$$J_{\sigma_i}(f_j) = \frac{\partial f_j}{\partial \sigma_i} = u_i u_i^T x \delta_{ij}$$

The δ_{ij} is added since σ_i only exists where $i = j$, and otherwise $\sigma = 0$. From here We get in matrix notation

$$J_\sigma(f) = \sum_{i=1}^{n} u_i u_i^T x \delta_{ij} = U \cdot diag(U^T x)$$

9. $h(\sigma) = \frac{1}{2}||f(\sigma) - y||^2$,

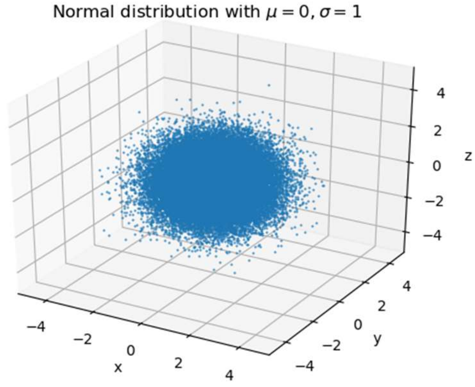
$$\nabla h_i = \frac{\partial h}{\partial \sigma_i} = \frac{\partial h}{\partial f}\frac{\partial f}{\partial \sigma_i} \to \nabla h = (f(\sigma) - y)^T \nabla f$$

10. $g(z)_j = e^{z_j} / \sum_{k=1}^{K} e^{z_k}$

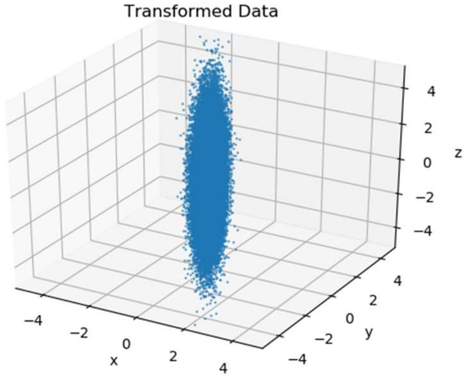
$$\big(J_z(g)\big)_{ij} = \frac{\partial g(z)_i}{\partial z_j} = \frac{\partial}{\partial z_j}\left(\frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}\right) = \frac{\frac{\partial e^{z_i}}{\partial z_j}\sum_{k=1}^{K} e^{z_k} + e^{z_i}\frac{\partial}{\partial z_j}\left(\sum_{k=1}^{K} e^{z_k}\right)}{\left(\sum_{k=1}^{K} e^{z_k}\right)^2}$$

$$= \frac{\delta_{ij} e^{z_i}\sum_{k=1}^{K} e^{z_k} + e^{z_i} e^{z_j}}{\left(\sum_{k=1}^{K} e^{z_k}\right)^2} = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}\left(\delta_{ij} + \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}\right) = g_i(z)\left(\delta_{ij} + g_j(z)\right)$$

PART II



11.

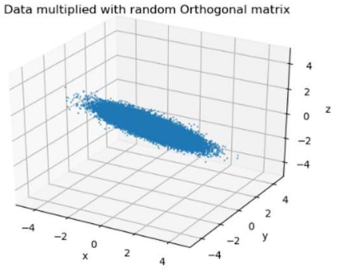


12.

The covariance matrix is

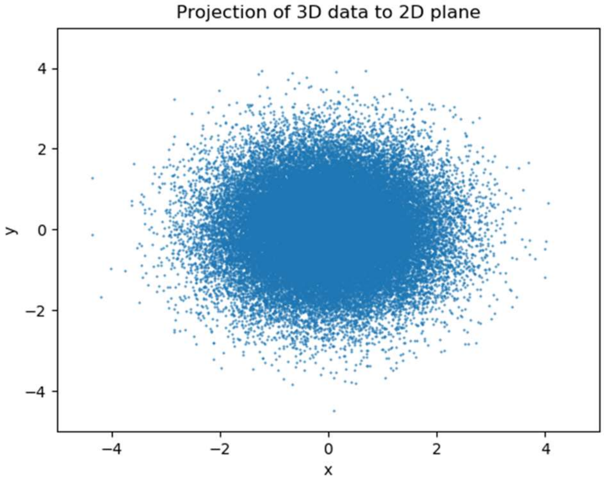
$$\begin{matrix}[[\ 1.00464665e-02 & 2.37037835e-04 & -1.58898654e-03] \\ [\ 2.37037835e-04 & 2.49184659e-01 & 3.89948457e-03] \\ [-1.58898654e-03 & 3.89948457e-03 & 4.00151813e+00]]\end{matrix}$$

Which is approximately S^2 (up to values that are close to zero outside the diagonal line). this was expected because the elements on the diagonal are the variance of the new data, and the variance of the new data (call it $\vec{r}$) is $Var(S\vec{r}) = S^2 Var(\vec{r})$.



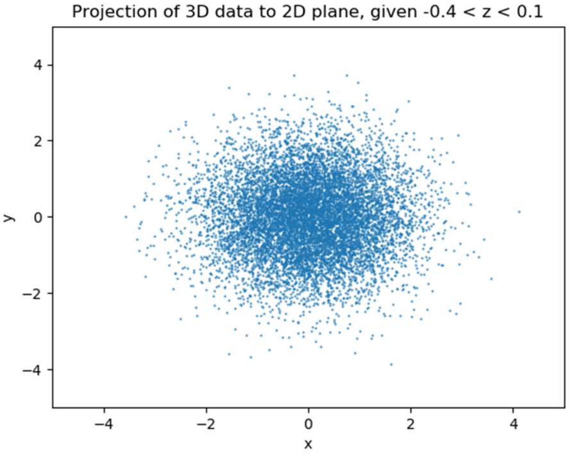
13.

After multiplication with a random unitary matrix, the covariance matrix isn't diagonal anymore, and contains random values outside the diagonal line.



14.

We can see that the projection yields a 2D gaussian distribution, as expected.

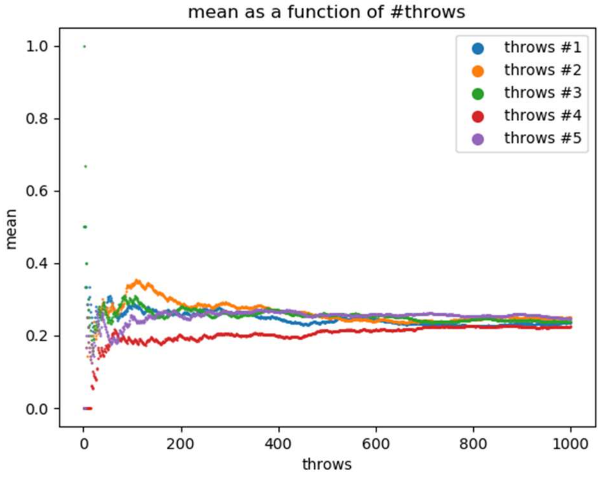


15.

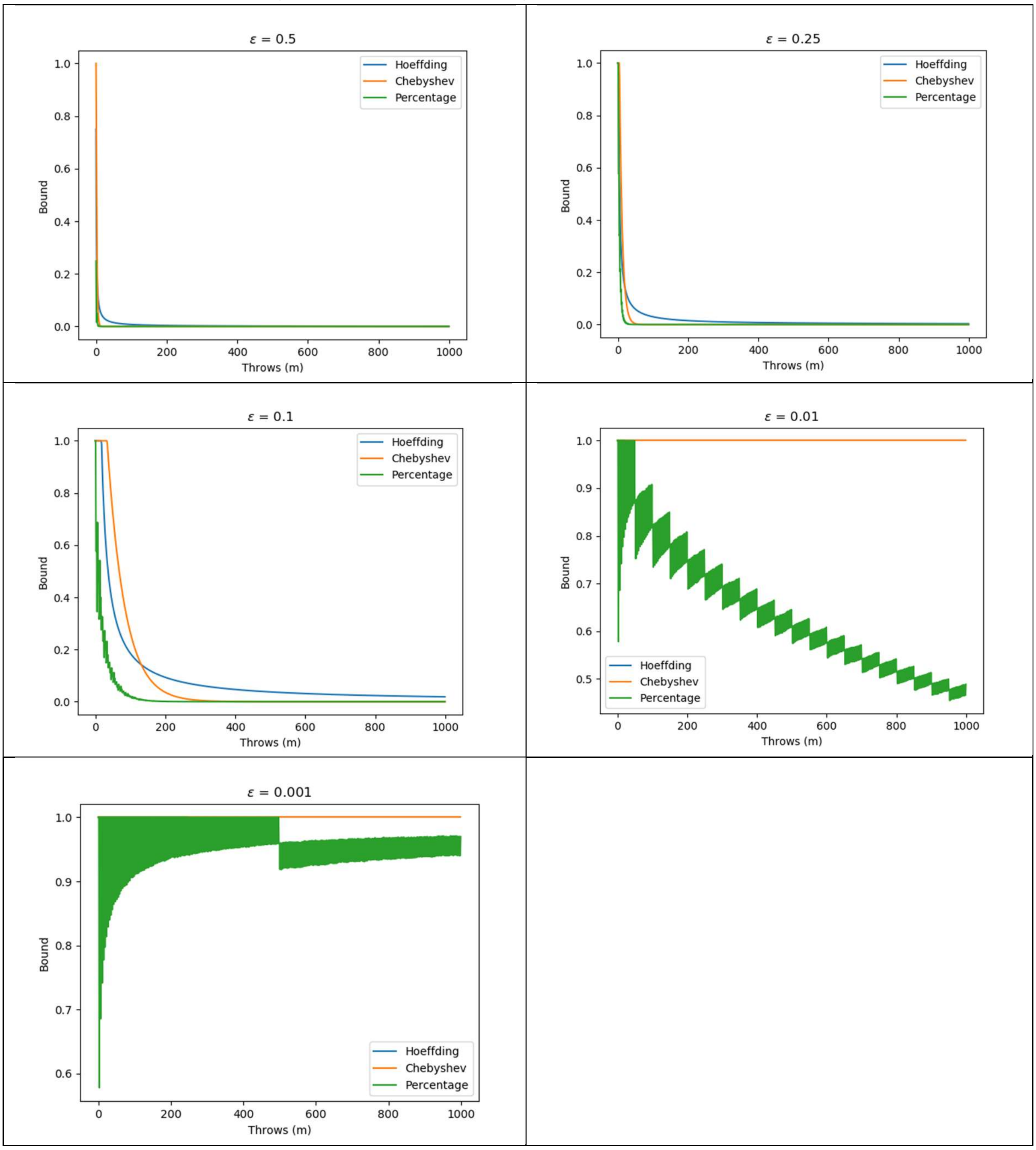
We can see that the projection remains a 2D gaussian (with less data points, due to the condition $-0.4 < z < 0{,}1$)

16.

A. we expect that the $mean$ as a function of $\#throws$ will approach constant value 0.25. this is due to the fact that a coin toss follows a binomial distribution, and in our case $n = 1000, p = 0.25$, therefore $X \sim Bin(1000{,}0.25) = 0.25$



B&C. the results for $\varepsilon \in \{0.5, 0.25, 0.1, 0.01, 0.001\}$:



For large values of ε: The *Percentage* plot has small values $\forall m$

For medium values of ε: The *Percentage* plot has significant values where m is small, and vice versa

For small values of ε: The *Percentage* plot has significant values $\forall m$. This is due to the fact values very close to the expectancy are scarce.