Part 1: Theory

1. X,Y are two random variables.

   a. The covariance of X,Y is:

   $$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

   Where E[X] and E[Y] are the expected values of X,Y respectively.
   By using the linearity property of expectation, and because $E\big[E[X]\big] = E[X]$:

   $$cov(X,Y) = \sigma_{X,Y} = E[(X - E[X])(Y - E[Y])] = E\big[XY - XE[Y] - YE[X] + E[X]E[Y]\big] =$$
   $$= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y]$$

   If X and Y are statistically independent, then $E[XY] = E[X]E[Y]$, so $cov(X,Y) = 0$
   If the covariance is equal to zero, then the correlation is equal to zero because:

   $$corr(X,Y) = \rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \xRightarrow{cov(X,Y)=0} corr(X,Y) = 0$$

   b. If X and Y are not correlated, i.e., $corr(X,Y) = 0$, we can not deduce that X,Y are independent.
   That's because there are pairs of dependent random variables that satisfies $cov(X,Y) = 0$.
   For example:
   X is uniformly distributed in $[-1,1]$,
   Y satisfies $Y = X^2$,
   X,Y are not independent (We will show why in section 2.a)
   The covariance of X,Y is:

   $$cov(X,Y) = E[XY] - E[X]E[Y] = E[X^3] - E[X]E[X^2]$$

   But:

   $$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-1}^{1} x \cdot \frac{1}{1-(-1)}\,dx = \frac{1}{2} \cdot \frac{x^2}{2}\Big|_{-1}^{1} = \frac{1-(-1)^2}{2} = 0$$
   $$E[X^3] = \int_{-\infty}^{\infty} x^3 f_X(x)\,dx = \int_{-1}^{1} x^3 \cdot \frac{1}{1-(-1)}\,dx = \frac{1}{2} \cdot \frac{x^4}{4}\Big|_{-1}^{1} = \frac{1^4-(-1)^4}{8} = 0$$

   $\Rightarrow cov(X,Y) = 0 \Rightarrow corr(X,Y) = 0 \Rightarrow X, Y$ are not correlated and not independent.

2. Let X be a random variable.

   a. X and X^2 are always independent.
   False
   Counter example:
   Let X be a uniformly random variable in the range [-1, 1], and $Y = X^2$.
   If Y=1, we can say that X must be either 1 or -1. Meaning that by knowing something on Y we now know something definite on X, and therefore they are **dependent**.

b. X and X^2 are never correlated.

False

Let X be a random variable with the following pmf:

$$X = \begin{cases} 2 & P = 0.5 \\ 4 & P = 0.5 \end{cases}$$

$E(X) = 2 \cdot 0.5 + 4 \cdot 0.5 = 3$

$E(X^2) = 2^2 \cdot 0.5 + 4^2 \cdot 0.5 = 10$

$E(X^3) = 2^3 \cdot 0.5 + 4^3 \cdot 0.5 = 36$

$\Rightarrow Cov(X, X^2) = E(X \cdot X^2) - E(X) \cdot E(X^2) = 36 - 30 \neq 0 \Rightarrow corr(X, X^2) \neq 0$

c. X and X^2 are always correlated.

False

The example from section 1.b can serve as a counter example for this statement:

X is uniformly distributed in $[-1,1]$,

Y satisfies $Y = X^2$,

The covariance of X, Y is:

$$cov(X, Y) = E[XY] - E[X]E[Y] = E[X^3] - E[X]E[X^2]$$

But:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-1}^{1} x \cdot \frac{1}{1-(-1)}\, dx = \frac{1}{2} \cdot \frac{x^2}{2}\Big|_{-1}^{1} = \frac{1-(-1)^2}{2} = 0$$

$$E[X^3] = \int_{-\infty}^{\infty} x^3 f_X(x)\, dx = \int_{-1}^{1} x^3 \cdot \frac{1}{1-(-1)}\, dx = \frac{1}{2} \cdot \frac{x^4}{4}\Big|_{-1}^{1} = \frac{1^4-(-1)^4}{8} = 0$$

$\Rightarrow cov(X, Y) = 0 \Rightarrow corr(X, Y) = 0 \Rightarrow X, Y$ are not correlated.

d. X and X^2 are never independent.

False

Two random variables are independent iff $P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} = \frac{P_X(x) \cdot P_Y(y)}{P_Y(y)} = P_X(x)$.

Let X be a random variable with the following pmf:

$$X = \{1\ with\ probability\ P = 1\}$$

Therefore, $X^2 = Y = 1$ with probability $P = 1$.

$\Rightarrow P_{X|Y}(x = 1|y = a) = 1, P_{X|Y}(x \neq 1|y = a) = 0 \Rightarrow P_{X|Y}(x|y) = P_X(x)$

$\Rightarrow X, Y\ are\ independent.$

3. Let X be a 4x4 matrix and $C = X^T X$.

$$C^T = (X^T X)^T = X^T (X^T)^T = X^T X = C$$

$\Rightarrow C^T = C \Rightarrow$ C is a symmetric matrix.

An 4x4 symmetric real matrix C is said to be positive-semidefinite if $z^T C z \geq 0$ for all $z \in \mathbb{R}^4$.

$$z^T C z = z^T X^T X z = (Xz)^T Xz$$

$(Xz)^T$ is a 1x4 vector and Xz is a 4x1 vector. $(Xz)^T Xz$ is the dot product $Xz \cdot Xz$ which is the square of the Euclidean norm $\|Xz\|_2 = \sqrt{(Xz)^2}$.

The Euclidian norm is always non-negative so:

$$\|Xz\|_2 \geq 0 \Rightarrow z^T C z \geq 0 \; \forall z \in \mathbb{R}^4 \Rightarrow C \; is \; positive - semidefinite$$

4. Matrix $X \in \mathbb{R}^{6\times 2}$ is given by:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 5 & 6 \\ 6 & 6 \\ 7 & 6 \end{bmatrix} \quad i.e. x_1 = [1 \quad 0], x_2 = [2 \quad 0], etc.$$

Where every row is an observation, and every column is a feature.

a. Before calculating the covariance matrix of X, we will center the data by subtracting the mean of each feature, so:

$$X_{center} \rightarrow \begin{bmatrix} 1-4 & 0-3 \\ 2-4 & 0-3 \\ 3-4 & 0-3 \\ 5-4 & 6-3 \\ 6-4 & 6-3 \\ 7-4 & 6-3 \end{bmatrix} = \begin{bmatrix} -3 & -3 \\ -2 & -3 \\ -1 & -3 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{bmatrix}$$

The covariance matrix of X_{center} is:

$$C_X = \underbrace{\frac{1}{6} X_{center}^T X_{center}}_{*} = \frac{1}{6}\begin{bmatrix} -3 & -2 & -1 & 1 & 2 & 3 \\ -3 & -3 & -3 & 3 & 3 & 3 \end{bmatrix}\begin{bmatrix} -3 & -3 \\ -2 & -3 \\ -1 & -3 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{bmatrix} = \begin{bmatrix} \frac{14}{3} & 6 \\ 6 & 9 \end{bmatrix}$$

* This form corresponds to a data matrix where the number of rows represents the number of samples and the number of columns represents the number of features (in contrary to the form we saw in the tutorial).

b. C_X is a real symmetric matrix and thus, according to the spectral decomposition theorem, can be diagonalized by its eigen matrix P.
We will start with finding the eigenvalues and their associated eigenvectors.

$$\det(\lambda I - C_X) = \begin{vmatrix} \lambda - \frac{14}{3} & -6 \\ -6 & \lambda - 9 \end{vmatrix} = \left(\lambda - \frac{14}{3}\right)(\lambda - 9) - (-6)(-6) = \lambda^2 - \frac{41}{3}\lambda + 6 = 0$$

$$\Rightarrow \lambda_{1,2} = 13.21, 0.45$$

Eigenvectors v_1, v_2 associated with eigenvalues λ_1, λ_2 solve:

$$(\lambda I - C_X)v = 0$$

$$\begin{bmatrix} \lambda - \frac{14}{3} & -6 \\ -6 & \lambda - 9 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

For $\lambda_1 = 13.21$:

$$\begin{bmatrix} 8.55 & -6 \\ -6 & 4.21 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 8.55v_{11} - 6v_{12} \\ -6v_{11} + 4.21v_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{cases} 8.55v_{11} - 6v_{12} = 0 \\ -6v_{11} + 4.21v_{12} = 0 \end{cases} \Rightarrow \begin{cases} v_{11} = 0.7v_{12} \\ v_{11} = 0.7v_{12} \end{cases}$$

$$v_1 = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 1 \\ 1.43 \end{bmatrix}$$

For $\lambda_2 = 0.45$:

$$\begin{bmatrix} -4.21 & -6 \\ -6 & -8.55 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} -4.21v_{21} - 6v_{22} \\ -6v_{21} - 8.55v_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{cases} v_{21} = -1.43v_{22} \\ v_{21} = -1.43v_{22} \end{cases}$$

$$v_2 = \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ -0.7 \end{bmatrix}$$

The diagonalizing matrix P is:

$$P = \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1.43 & -0.7 \end{bmatrix}$$

We can now diagonalize C_X by calculating $P^{-1}C_X P$.

$$P^{-1} = \frac{1}{-0.7 - 1.43} \begin{bmatrix} -0.7 & -1 \\ -1.43 & 1 \end{bmatrix} = -0.47 \begin{bmatrix} -0.7 & -1 \\ -1.43 & 1 \end{bmatrix} = \begin{bmatrix} 0.33 & 0.47 \\ 0.67 & -0.47 \end{bmatrix}$$

$$C_{X,D} = \underbrace{\begin{bmatrix} -0.33 & -0.47 \\ -0.67 & 0.47 \end{bmatrix}}_{P^{-1}} \underbrace{\begin{bmatrix} \frac{14}{3} & 6 \\ 6 & 9 \end{bmatrix}}_{C_X} \underbrace{\begin{bmatrix} 1 & 1 \\ 1.43 & -0.7 \end{bmatrix}}_{P} = \begin{bmatrix} 13.21 & 0 \\ 0 & 0.45 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

c. The principal components are eigenvectors of the centered data's covariance matrix.
The variance is maximal for the largest eigenvalue, so the main component (PC1) corresponds to the eigenvalue $\lambda_1 = 13.21$ (the variance) and the eigenvector $v_1 = \begin{bmatrix} 1 \\ 1.43 \end{bmatrix}$ (the direction).
This component carry $\frac{13.21}{13.21+0.45} \cdot 100\% = 96.7\%$ of the variance of the data.

d. As said, the main eigenvalue is $\lambda_1 = 13.21$ because it corresponds to maximal variance in the data.

e. We will now calculate the projection of the new data point $x_7 = (2 \quad 1)$ on the main component.
First, we will normalize the eigenvector corresponds to the main component:

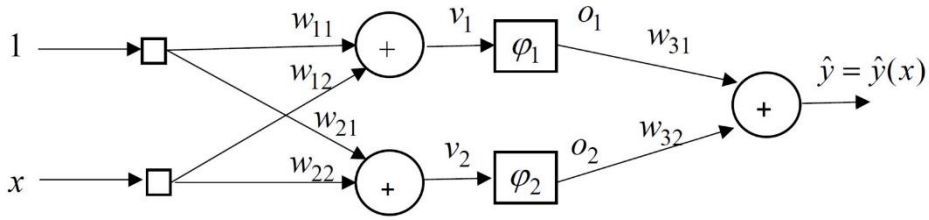
$$v_1^T v_1 = 1 \rightarrow \hat{v}_1 = \begin{bmatrix} 0.57 \\ 0.82 \end{bmatrix}$$

After normalizing x_7 by subtracting the mean from each feature, we can calculate the projection of x_7 on the main component as follows:

$$x_7 \to \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$$

$$p_{x_7} = [0.57 \quad 0.82] \begin{bmatrix} -2 \\ -2 \end{bmatrix} = -2.78$$

5. We are given with the following neural network:



a. For each weights matrix, $W^{(i)} \in \mathbb{R}^{n \times p}$ where n is the number of inputs before getting to layer number i, and p is the number of neurons in layer number i. Thus:

$$W^{(1)} \in \mathbb{R}^{2\times 2}, \; W^{(1)} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

$$W^{(2)} \in \mathbb{R}^{2\times 1}, \; W^{(2)} = \begin{bmatrix} w_{31} \\ w_{32} \end{bmatrix}$$

b.

$$X = \begin{bmatrix} 1 \\ x \end{bmatrix} \quad \varphi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix}$$

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = W^{(1)}X \quad o = \begin{bmatrix} o_1 \\ o_2 \end{bmatrix} = \varphi v$$

$$\hat{y}(x) = \underbrace{W^{T(2)}}_{\in \mathbb{R}^{1\times 2}} \underbrace{\left(\varphi \underbrace{\left(W^{(1)}X \right)}_{v \in \mathbb{R}^{2\times 1}} \right)}_{o} \in \mathbb{R}^{1\times 1}$$

Where:

$$W^{(1)}X = \begin{bmatrix} w_{11} \cdot 1 + w_{12} \cdot x \\ w_{21} \cdot 1 + w_{22} \cdot x \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

φ_1, φ_2 are ReLU, i.e., $\varphi_1(z) = \varphi_2(z) = \max\{0, z\}$, so:

$$\hat{y}(x) = [w_{31} \quad w_{32}] \cdot \begin{bmatrix} \max\{0, w_{11} + w_{12} \cdot x\} \\ \max\{0, w_{21} + w_{22} \cdot x\} \end{bmatrix}$$

$$\hat{y}(x) = w_{31} \cdot \max\{0, w_{11} + w_{12} \cdot x\} + w_{32} \cdot \max\{0, w_{21} + w_{22} \cdot x\}$$

c. $y(x)$ can be explicitly written as follows:

$$y(x)=\begin{cases}-\frac{1}{2}(x+1) & x<-1\\ 0 & -1\le x\le 1\\ x-1 & x>1\end{cases}=\max\left\{0,\underbrace{\left(\begin{cases}-\frac{1}{2}(x+1) & x<0\\ x-1 & x\ge 0\end{cases}\right)}_{u(x)}\right\}=\varphi(\mathrm{u(x)})$$

If:

$$W^{(1)}=\begin{bmatrix}w_{11} & w_{12}\\ w_{21} & w_{22}\end{bmatrix}=\begin{bmatrix}-\frac{1}{2} & -\frac{1}{2}\\ -1 & 1\end{bmatrix}$$

$$W^{(2)}=\begin{bmatrix}w_{31}\\ w_{32}\end{bmatrix}=\begin{bmatrix}1\\ 1\end{bmatrix}$$

So:

$$\hat{y}(x)=[w_{31}\quad w_{32}]\cdot\begin{bmatrix}\max\{0,w_{11}+w_{12}\cdot x\}\\ \max\{0,w_{21}+w_{22}\cdot x\}\end{bmatrix}=[1\quad 1]\cdot\begin{bmatrix}\max\left\{0,-\frac{1}{2}-\frac{1}{2}x\right\}\\ \max\{0,-1+x\}\end{bmatrix}$$

And then:

$$\hat{y}(x)=\begin{cases}[1\quad 1]\cdot\begin{bmatrix}-\frac{1}{2}(x+1)\\ 0\end{bmatrix} & x<-1\\ [1\quad 1]\cdot\begin{bmatrix}0\\ 0\end{bmatrix} & -1\le x\le 1\\ [1\quad 1]\cdot\begin{bmatrix}0\\ x-1\end{bmatrix} & x>1\end{cases}=\begin{cases}-\frac{1}{2}(x+1) & x<-1\\ 0 & -1\le x\le 1\\ x-1 & x>1\end{cases}=y(x)$$

d. The loss is given as follows:

$$L(y,\hat{y})=\left(y-\hat{y}(x)\right)^2$$

$$y(x)=\max\{0,u(x)\}$$

$$\hat{y}(x)=W^{T(2)}\underbrace{\left(\varphi\underbrace{\left(W^{(1)}X\right)}_{v}\right)}_{o}$$

$$\hat{y}(x)=w_{31}\cdot\max\{0,w_{11}+w_{12}\cdot x\}+w_{32}\cdot\max\{0,w_{21}+w_{22}\cdot x\}$$

Our model's parameters are $W^{(1)},W^{(2)}$ so we need to derive L according to $W^{(1)},W^{(2)}$ in order to use it for gradient descent.

$$\vec{\nabla}L=\left(\frac{\partial L}{\partial W^{(1)}},\frac{\partial L}{\partial W^{(2)}}\right)$$

$$\frac{\partial L}{\partial W^{(1)}}=\frac{\partial L}{\partial\hat{y}}\cdot\frac{\partial\hat{y}}{\partial o}\cdot\frac{\partial o}{\partial v}\cdot\frac{\partial v}{\partial W^{(1)}}=-2\left(y-\hat{y}(x)\right)\cdot W^{T(2)}\cdot ReLU'(v)\cdot X$$

Where $ReLU'(v) = \begin{cases} 0 & v < 0 \\ 1 & v \geq 0 \end{cases}, v = W^{(1)}X.$

$$\frac{\partial L}{\partial W^{(2)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial W^{(2)}} = -2\big(y - \hat{y}(x)\big) \cdot o^T$$

Where $o = \varphi\big(W^{(1)}X\big), \varphi\ is\ the\ ReLU\ activation\ function.$

e. The update policy of every weight is done by the gradient descent process:

$$W_{n+1}^{(j)} = W_n^{(j)} - \eta \cdot \frac{\partial L}{\partial W_n^{(j)}}$$

where $j \in [1,2]$, η is the learning rate.

f. For the following weights initialization $W^{(1)} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$, $W^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
We can calculate the first value of the loss for the following sample $(x, y) = (0,0)$ as follows:
First, we'll calculate the predicted y value ($\hat{y}$):

$$\Rightarrow \hat{y}(x) = W^{T(2)}\Big(\varphi\big(W^{(1)}X\big)\Big) = [1 \quad 1] \cdot \begin{bmatrix} \max\{0, 1+x\} \\ \max\{0, -(1+x)\} \end{bmatrix}$$

$$\Rightarrow \hat{y}(0) = [1 \quad 1] \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1$$

The loss function is $L(y, \hat{y}) = \big(y - \hat{y}(x)\big)^2$, so:

$$L(0,1) = (0-1)^2 = 1$$