Name: SOLUTION **Student ID:** SOLUTION

If you "submit images" on Gradescope, leave name and UniMelb ID blank. If you "submit PDF", write name and student ID.

Instructions

- This homework is **not for credit**, based on a total of 10 points.
- Due on July 28, 11.59pm AEST.
- In all the questions, you will get the points only when the whole answer is correct. (There will not be any partial points.)

Questions

1) [1 point] Consider the following joint probabilities:

$P(X = 1, Y = 1) = 9/16$ $\quad P(X = 2, Y = 1) = 3/16$

$P(X = 1, Y = 2) = 3/16$ $\quad P(X = 2, Y = 2) = 1/16$

Are X and Y independent?

Answer "yes" or "no", and support your answer with some computations.

Yes, X and Y are independent.

$P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = 2) = 9/16 + 3/16 = 12/16 = 3/4$
$P(X = 2) = P(X = 2, Y = 1) + P(X = 2, Y = 2) = 3/16 + 1/16 = 4/16 = 1/4$
$P(Y = 1) = P(X = 1, Y = 1) + P(X = 2, Y = 1) = 9/16 + 3/16 = 12/16 = 3/4$
$P(Y = 2) = P(X = 1, Y = 2) + P(X = 2, Y = 2) = 3/16 + 1/16 = 4/16 = 1/4$

$P(X = 1, Y = 1) = P(X = 1)P(Y = 1)$?, or equivalently, $9/16 = 3/4 \times 3/4$? Yes
$P(X = 1, Y = 2) = P(X = 1)P(Y = 2)$?, or equivalently, $3/16 = 3/4 \times 1/4$? Yes
$P(X = 2, Y = 1) = P(X = 2)P(Y = 1)$?, or equivalently, $3/16 = 1/4 \times 3/4$? Yes
$P(X = 2, Y = 2) = P(X = 2)P(Y = 2)$?, or equivalently, $1/16 = 1/4 \times 1/4$? Yes

2) [1 point] Let $\mathbf{A} = \begin{bmatrix} 3 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 4 \end{bmatrix}$ and $\mathbf{b} = (7, 6, 5)^{\mathrm{T}}$. Let $\mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}}$.

Can we express $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x}$ in the following fashion?

$$f(\mathbf{x}) = c_1x_1^2 + c_2x_2^2 + c_3x_3^2 + c_4x_1x_2 + c_5x_1x_3 + c_6x_2x_3 + c_7x_1 + c_8x_2 + c_9x_3$$

If so, what are the values of $c_1, \ldots, c_9$?

$c_1 = c_2 = 3/2$, $c_3 = 2$, $c_4 = -2$, $c_5 = 0$, $c_6 = -1$, $c_7 = 7$, $c_8 = 6$, $c_9 = 5$.

Name: SOLUTION **Student ID:** SOLUTION

If you "submit images" on Gradescope, leave name and UniMelb ID blank. If you "submit PDF", write name and student ID.

3) [1 point] Let $\mathbf{x} = (x_1, x_2, x_3)$ and let $f(\mathbf{x}) = x_1 e^{x_2} + 3e^{x_3}$. Compute the gradient $\nabla f(\mathbf{x}) \in \mathbb{R}^3$.

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix} = \begin{bmatrix} e^{x_2} \\ x_1 e^{x_2} \\ 3e^{x_3} \end{bmatrix}$$

4) [0.5 points] Let $\mathbf{x} = (x_1, x_2, x_3)$ and let $f(\mathbf{x}) = x_1 \sin(x_2) + \cos(x_3)$. Compute the Hessian $\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{3\times 3}$.

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 x_1} & \frac{\partial^2 f}{\partial x_1 x_2} & \frac{\partial^2 f}{\partial x_1 x_3} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2 x_2} & \frac{\partial^2 f}{\partial x_2 x_3} \\ \frac{\partial^2 f}{\partial x_3 x_1} & \frac{\partial^2 f}{\partial x_3 x_2} & \frac{\partial^2 f}{\partial x_3 x_3} \end{bmatrix} = \begin{bmatrix} 0 & \cos(x_2) & 0 \\ \cos(x_2) & -x_1 \sin(x_2) & 0 \\ 0 & 0 & -\cos(x_3) \end{bmatrix}$$

5) [1 point] Let $\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$. What is $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$, $(\mathbf{AB})^{-1}$, $\det \mathbf{A}$, $\det \mathbf{B}$ and $\det(\mathbf{AB})$?

$\mathbf{A}^{-1} = \frac{1}{5}\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$, $\mathbf{B}^{-1}$ and $(\mathbf{AB})^{-1}$ are undetermined.
$\det(\mathbf{A}) = 5$, $\det(\mathbf{B}) = \det(\mathbf{AB}) = 0$

6) [1 point] Let $\mathbf{A} = \begin{bmatrix} 1 & -\varepsilon \\ -\varepsilon & 1 \end{bmatrix}$. Compute $\mathbf{A}^2$.

$$\mathbf{A}^2 = \mathbf{AA} = \begin{bmatrix} 1+\varepsilon^2 & -2\varepsilon \\ -2\varepsilon & 1+\varepsilon^2 \end{bmatrix}$$

7.a) [1 point] Let y be a random variable with value -1 with probability 0.4, with value 0 with probability 0.15, and with value $+1$ with probability 0.45. What is the expected value of y?

$$\mathbb{E}[y] = -1 \times \mathbb{P}[y=-1] + 0 \times \mathbb{P}[y=0] + 1 \times \mathbb{P}[y=+1] = -1 \times 0.4 + 1 \times 0.45 = 0.05$$

Name: SOLUTION **Student ID:** SOLUTION

If you "submit images" on Gradescope, leave name and UniMelb ID blank. If you "submit PDF", write name and student ID.

7.b) [1 point] Write a Python function that generates T random values of y (defined as in 7.a), and then computes the average of those T values. For the purpose of grading, please provide the function in writing. (We will not run your code, although the function should be correct.)

```
import numpy as np
def expval(T):
   p = np.random.random(T)
   y = np.zeros(T)
   y[p<0.4] = -1
   y[p>0.55] = 1
   return np.mean(y)
```

7.c) [0.5 points] Given your Python function (from 7.b) report the averages you found for $T = 40$, $T = 400$ and $T = 40000$. Is the average closer to or further than the expected value of y (found in 7.a) for larger values of T? You have to run your program several times in order to see a clear pattern.

`abs(expval(40)) > abs(expval(400)) > abs(expval(40000))` very often. For instance for one run, `expval(40)=0.225`, `expval(400)=0.0675` and `expval(40000)=0.0512`. Thus, the average is closer to the expected value of y for larger values of T.

8) [1 point] Let y be a Gaussian random variable with mean 0 and variance 1, and let z be some *continuous* random variable (not necessarily Gaussian). If the covariance between y and z is 0, does that mean that y and z are independent? Give a specific example/counterexample that describes z or the probability distribution of z precisely.

Let $z = y^2$. Clearly y and z are not independent. We know that $\mathbb{E}[y] = 0$, $\mathbb{E}[y^2] = 1$, $\mathbb{E}[y^3] = 0$ and therefore:
$\mathbb{C}\text{ov}(y, z) = \mathbb{E}[yz] - \mathbb{E}[y] \times \mathbb{E}[z] = \mathbb{E}[y^3] - \mathbb{E}[y] \times \mathbb{E}[y^2] = 0 - 0 \times 1 = 0$

Name: SOLUTION **Student ID:** SOLUTION

If you "submit images" on Gradescope, leave name and UniMelb ID blank. If you "submit PDF", write name and student ID.

9) [1 point] Bob and Mary have a classification dataset of 100 samples and 300 features. Bob and Mary decide to use two-fold cross validation with principal component analysis (PCA), to reduce the 300 features to only 3 features.
Bob has the following implementation:

- The input data is $X \in \mathbb{R}^{100\times300}$ and $y \in \{-1,+1\}^{100}$
- Use PCA to find the 3 orthogonal directions $U \in \mathbb{R}^{300\times3}$ of largest variance from X (after centering) and store the projected data in $Z \in \mathbb{R}^{100\times3}$
- Split Z and y in two equal-size sets $Z' \in \mathbb{R}^{50\times3}$, $y' \in \{-1,+1\}^{50}$, and $Z'' \in \mathbb{R}^{50\times3}$, $y'' \in \{-1,+1\}^{50}$
- Train on Z', y', and test on Z'', y''
- Train on Z'', y'', and test on Z', y'

Mary has the following implementation:

- The input data is $X \in \mathbb{R}^{100\times300}$ and $y \in \{-1,+1\}^{100}$
- Split X and y in two equal-size sets $X' \in \mathbb{R}^{50\times300}$, $y' \in \{-1,+1\}^{50}$, and $X'' \in \mathbb{R}^{50\times300}$, $y'' \in \{-1,+1\}^{50}$
- Use PCA to find the 3 orthogonal directions $U' \in \mathbb{R}^{300\times3}$ of largest variance from X' (after centering) and store the projected data in $Z' \in \mathbb{R}^{50\times3}$
- Use the same 3 orthogonal directions U' and centering (from the previous step) to transform X'' and store the projected data in $Z'' \in \mathbb{R}^{50\times3}$
- Train on Z', y', and test on Z'', y''
- Use PCA to find the 3 orthogonal directions $U'' \in \mathbb{R}^{300\times3}$ of largest variance from X'' (after centering) and store the projected data in $Z'' \in \mathbb{R}^{50\times3}$
- Use the same 3 orthogonal directions U'' and centering (from the previous step) to transform X' and store the projected data in $Z' \in \mathbb{R}^{50\times3}$
- Train on Z'', y'', and test on Z', y'

Who has the correct implementation? Explain your answer.

In Bob's implementation, there are some issues.

When training on Z', y', and testing on Z'', y'':
Z' depends on Z, and Z depends on X, thus on all the samples (training and testing).
Therefore, when training, the algorithm can "see" the test data.

When training on Z'', y'', and testing on Z', y':
Z'' depends on Z, and Z depends on X, thus on all the samples (training and testing).
Therefore, when training, the algorithm can "see" the test data.

Since Mary splits X into two sets X' and X'', this problem does not occur in her implementation.