# Problem Set 1

**Oke**

CEE 616: Probabilistic Machine Learning *09.11.2025*

*Due Thursday, September 25, 2025 at 11:59PM on Canvas.*

The problems are worth a total of **85 points**.

### Programming requirements

It is recommended that you have a working installation of JupyterLab. Google Colab is also another viable notebook option. However, notebooks are optional, and you can always choose to write your code in an editor (e.g. Sublime Text) and process your outputs accordingly.

In Jupyter, you may use a Python, R or even MATLAB kernel.[1] Text can be formatted using Markdown (brief guide here: `https://learn.getgrav.org/content/markdown`).

Any Python packages you require must be installed locally to access them via the Jupyter kernel. (Use `conda` or `pip`.)

### Programming help

I will provide some programming templates in Python/R in the coming days on Moodle, in order to ease some of the possible issues that may arise as you begin scripting. We will also go over coding examples weekly.

### Submission instructions

There are two options for submission:

1. JupyterLab Notebook. Please name your notebook as follows:
   `<lastname>-<firstname>-PS1.ipynb`

2. R/Python/MATLAB script *and* PDF document with supporting responses. Your PDF should have complete responses to all the questions (including all the required plots). Your script should be clearly commented, producing all the results and plots you show in your PDF document. The filenames should be in a similar format as described above.

[1] In order to use the R kernel in Jupyter, you may have to install it, as well. Please follow the brief instructions here: `https://irkernel.github.io/installation/`. To install the MATLAB kernel, see `https://am111.readthedocs.io/en/latest/jmatlab_install.html`.

## Problem 1 *True/False questions (10 points)*

Respond "T" (*True*) or "F" (*False*) to the following statements. Use the boxes provided. Each response is worth 1 point.

**(i)** **F** Supervised learning only constitutes modeling frameworks that are suitable for regression. Supervised learning includes both regression and classification.

**(ii)** **F** Bayes' Theorem specifies a posterior probability $P(X|A)$ as inversely proportional to the product of its likelihood $P(A|X)$ and prior $P(X)$. Directly proportional, not inversely

**(iii)** **T** In the gradient descent optimization method, the second derivative of the loss function is not required to compute the update step.

**(iv)** **F** Uncorrelated random variables are always independent of each other. Correlation only measures linear dependency.

**(v)** **F** In assessing a classifier, the receiver operating characteristics (ROC) curve is generated by plotting sensitivity versus specificity for different threshold values. versus $1-$ specificity

**(vi)** **T** The determinant of a matrix can be given as the product of its eigenvalues.

**(vii)** **F** The inverse of a matrix only exists if the matrix is singular. nonsingular

**(viii)** **F** An estimator is biased if the difference between its expectation and true value is zero. unbiased

**(ix)** **T** Leave-one-out-cross-validation (LOOCV) can be considered a special case of $k$-fold CV in which $k = n$, where $n$ is the number of observations in the dataset.

**(x)** **T** In bootstrapping, each observation $x_i$ in the original dataset is equally likely to be randomly selected for each resampled point in the sample.

## Problem 2 *Bayes' theorem (9 pts)*

Given that $P(A) = 0.6$, $P(B) = 0.3$ and $P(C) = 0.1$ represent the production of machines in a factory. The conditional probabilities of defective items are $P(D|A) = 0.02$, $P(D|B) = 0.03$ and $P(D|C) = 0.04$.

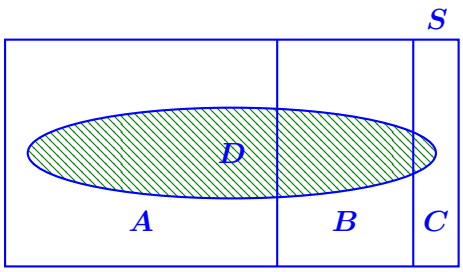
**(a)** Find the probability $P(D)$. We apply the theorem of total probability. *[3]*

$$P(D) = 0.02(0.6) + 0.03(0.3) + 0.04(0.1) = \boxed{0.025}$$

**(b)** Find the probability that an item was produced by machine A, given that it is defective. *[3]*

$$\text{Bayes' Theorem: } P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{0.02(0.6)}{0.025} = \boxed{0.48}$$

**(c)** Draw a Venn diagram depicting the interaction among the events $A$, $B$, $C$ and $D$ in sample space $S$. *[3]*




## Problem 3 *Estimating a linear model (16 pts)*

Given a vector of predictor variable samples $\boldsymbol{x}^T = \begin{bmatrix}10 & 5 & 7 & 19 & 11 & 8\end{bmatrix}$ and corresponding response vector $\boldsymbol{y}^T = \begin{bmatrix}15 & 9 & 3 & 25 & 7 & 13\end{bmatrix}$, the classical linear regression model can be written as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}$$

**(a)** Write the design matrix $\boldsymbol{X}$ in full assuming the model has an intercept (hint: the matrix should have two columns, the first being a column of 1's). *[2]*

$$\boldsymbol{X} = \begin{bmatrix}1 & 10\\ 1 & 5\\ 1 & 7\\ 1 & 19\\ 1 & 11\\ 1 & 8\end{bmatrix}$$

**(b)** Using OLS (ordinary least squares) assumptions (see equation 4.59 on page 113 in **PMLI**), find $\hat{\boldsymbol{w}}$. (Show your work as much as possible, but the matrix multiplication can be done in Python/R/MATLAB. Include the code used if you are not submitting a Jupyter notebook.) *[4]*

Given all ordinary least squares (OLS) assumptions hold, the expression for the best linear unbiased estimator, $\hat{\boldsymbol{\beta}}$ is defined in matrix form as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

In R, for example, the code to compute the estimates might look like this:

```
library(matlib)
z_mat = matrix(data = c(1,1,1,1,1,1,10,5,7,19,11,8),nrow=6,ncol=2)
y = matrix(data = c(15, 9, 3, 25, 7, 13),nrow=6)
z_1 = matrix(data = c(10, 5, 7, 19, 11, 8), nrow=6)
hatbeta = inv(t(z_mat)%*%z_mat)%*%(t(z_mat))%*%y
y_hat = z_mat%*%hatbeta
```

The solution is:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.67 \\ 1.27 \end{bmatrix}$$

*[2pts]* **(c)** Find the vector of predicted values $\hat{\boldsymbol{y}}$.

From the above code, we obtain:

$$\hat{y}^T = \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T = \begin{bmatrix} 12 & 5.67 & 8.2 & 23.4 & 13.67 & 9.47 \end{bmatrix}$$

*[2pts]* **(d)** Find the vector of residuals $\boldsymbol{e}$.

```
res = y - y_hat
```

The residual vector $\boldsymbol{e}$ is given by:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} 3.00 \\ 3.33 \\ -5.20 \\ 1.60 \\ -6.27 \\ 3.53 \end{bmatrix}$$

*[2pts]* **(e)** Compute the mean squared error (MSE) of this model.

In R, we use:

```
mse = (1/len(res))*sum(res^2)
```

The MSE is given by:

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \\ &= \frac{1}{6}\,(101.47) \\ &= \boxed{16.91} \end{aligned}$$

**(f)** Compute the root mean squared error (RMSE) of the model. *[1]*

$$\begin{aligned} RMSE &= \sqrt{MSE} \\ &= \sqrt{16.91} \\ &= \boxed{4.11} \end{aligned}$$

**(g)** Create a scatterplot of the data and show the least squares line in the plot. *[3]*

## Problem 4 *Logistic function (6 pts)*

The logistic sigmoid function is given by:

$$\boldsymbol{\sigma}(z) := \frac{1}{1+e^{-z}} \tag{1}$$

**(a)** Produce a plot (in Python/R/Jupyter) of the function in the domain $z \in [-5, 5]$. *[2]*

See `PS1-Solution.ipynb`.

**(b)** Show that its derivative is given by: *[4]*

$$\boldsymbol{\sigma}'(z) = \frac{d\boldsymbol{\sigma}(z)}{dz} = [1 - \boldsymbol{\sigma}(z)]\boldsymbol{\sigma}(z) \tag{2}$$

We use the chain rule:

$$\begin{aligned} \boldsymbol{\sigma}'(z) &= -1(-e^{-z})\frac{1}{(1+e^{-z})^2} \\ &= \frac{e^{-z}}{(1+e^{-z})(1+e^{-z})} \\ &= \left[\frac{e^{-z}}{1+e^{-z}}\right]\frac{1}{1+e^{-z}} \\ &= \left[\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\right]\boldsymbol{\sigma}(z) \\ &= [1 - \boldsymbol{\sigma}(z)]\,\boldsymbol{\sigma}(z) \end{aligned}$$

## Problem 5 *Classifier performance (14 points)*

An estimated classification model produces the following confusion matrix on a test set:

| | | *Predicted* | | |
|---|---|---|---|---|
| | | Class 0 | Class 1 | Total |
| *Observed* | Class 0 | 9650 | 17 | ? |
| | Class 1 | 265 | 68 | ? |
| | Total | ? | ? | |

*[1]* **(a)** What is the number of false positives ($FP$)? $FP = 17$

*[1]* **(b)** What is the number of false negatives ($FN$)? $FN = 265$

*[1]* **(c)** What is the number of positive observations ($P$)? $P = FN + TP = 265 + 68 = 333$

*[1]* **(d)** What is the number of predicted positive observations ($P^*$)? $P^* = FP + TP = 17 + 68 = 85$

*[2]* **(e)** Compute the test precision of the classifier.

$$\mathcal{P} = \frac{TP}{TP + FP} = \frac{TP}{P^*} = \frac{68}{68 + 17} = \frac{68}{85} = \boxed{0.8}$$

*[2]* **(f)** Compute the test recall of the classifier.

$$\mathcal{R} = \frac{TP}{TP + FN} = \frac{TP}{P} = \frac{68}{68 + 265} = \frac{68}{333} = \boxed{0.2}$$

*[2]* **(g)** Compute the test $F_1$-score of the classifier.

$$F_1 = 2 \cdot \frac{\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}} = 2 \cdot \frac{0.8(0.2)}{0.8 + 0.2} = \boxed{0.32}$$

*[2]* **(h)** Compute the test accuracy of the classifier.

$$\mathcal{A} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{68 + 9650}{68 + 9650 + 17 + 265} = \frac{9718}{10000} = \boxed{0.97}$$

## Problem 6 *Entropy (12 pts)*

**PMLI** Exercise 6.3 (page 218).

*[2]* **(a)** The joint entropy is given by

$$\begin{aligned}\mathbb{H}(X,Y) &= -\sum_{x,y} p(x,y)\log_2 p(x,y)\\ &= -\left(3(0)\log_2 0 + \frac{1}{4}\log_2\frac{1}{4} + 2\frac{1}{8}\log_2\frac{1}{8} + 6\frac{1}{16}\log_2\frac{1}{16} + 4\frac{1}{32}\log_2\frac{1}{32}\right)\\ &= -\left(0 - \frac{2}{4} - \frac{3}{4} - \frac{6}{4} - \frac{5}{8}\right)\\ &= 3\tfrac{3}{8}\end{aligned}$$

**(b)** The marginal entropies $\mathbb{H}(X)$ and $\mathbb{H}(Y)$ are given as follows. *[4]*

First, we sum the columns to get $p(x)$ and we sum the rows to get $p(y)$.

$$\begin{aligned}p(X=1) &= \frac{1}{8}+\frac{1}{16}+\frac{1}{16}+\frac{1}{4} = \frac{1}{2}\\ p(X=2) &= \frac{1}{16}+\frac{1}{8}+\frac{1}{16}+0 = \frac{1}{4}\\ p(X=3) &= \frac{1}{32}+\frac{1}{32}+\frac{1}{16}+0 = \frac{1}{8}\\ p(X=4) &= \frac{1}{32}+\frac{1}{32}+\frac{1}{16}+0 = \frac{1}{8}\\ p(Y=1) &= \frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\frac{1}{32} = \frac{1}{4}\\ p(Y=2) &= \frac{1}{16}+\frac{1}{8}+\frac{1}{32}+\frac{1}{32} = \frac{1}{4}\\ p(Y=3) &= \frac{1}{16}+\frac{1}{16}+\frac{1}{16}+\frac{1}{16} = \frac{1}{4}\\ p(Y=4) &= \frac{1}{4}+0+0+0 = \frac{1}{4}\end{aligned}$$

Then:

$$\begin{aligned}\mathbb{H}(X) &= -\mathbb{E}[\log_2 p(X)] = -\sum_x p(x)\log_2 p(x)\\ &= -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + 2\frac{1}{8}\log_2\frac{1}{8}\right) = -\left(-\frac{1}{2}-\frac{1}{2}-\frac{3}{4}\right)\\ &= 1\tfrac{3}{4}\\ \mathbb{H}(Y) &= -\mathbb{E}[\log_2 p(Y)] = -\sum_y p(y)\log_2 p(y)\\ &= -\left(4\frac{1}{4}\log_2\frac{1}{4}\right) = 2\end{aligned}$$

*[4]* **(c)** The conditional entropies of $X$ on specific values $y$ are given by $\mathbb{H}(X|Y=y) = -\sum_x p(x|y)\log p(x|y)$. We evaluate them as follows:

First, we note that:

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

Thus, for $X = [1,2,3,4]$:

$$p(X|Y=1) = 4\left[\frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}\right] = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right]$$

$$\text{Similarly, } p(X|Y=2) = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right]$$

$$p(X|Y=3) = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right]$$

$$p(X|Y=4) = [1,0,0,0]$$

The conditional entropies are therefore:

$$\mathbb{H}(X|Y=1) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + 2\frac{1}{8}\log_2\frac{1}{8}\right)$$
$$= -\left(-\frac{1}{2} - \frac{1}{2} - \frac{3}{4}\right) = 1\tfrac{3}{4}$$
$$\mathbb{H}(X|Y=2) = 1\tfrac{3}{4}$$
$$\mathbb{H}(X|Y=3) = -\left(4\frac{1}{4}\log_2\frac{1}{4}\right) = 2$$
$$\mathbb{H}(X|Y=4) = -\left(\log_2 1 + 0 + 0 + 0\right) = 0$$

The prior entropy on $X$ is $\mathbb{H}(X) = 1\frac{3}{4}$. The posterior entropy on $X$, that is $\mathbb{H}(X|Y)$ increases when the conditional (posterior) variance of $X$ given a certain value of $Y$, $\mathbb{V}[X|Y=y]$ is greater than the prior (marginal) variance of $X$, $\mathbb{V}[X]$. For instance, at $Y=3$, each value of $X$ is equally probable. Thus, there is more spread in that posterior (conditional) distribution. Hence the entropy *increases* compared to the prior. For $Y=1$ and $Y=2$, the posterior distributions are the same as the prior. Thus, the entropy does not change. For $Y=4$, the posterior has the least variance compared to the other three, as the probability of $X=1$ given $Y=4$ is a certainty (100%). The posterior entropy is thus 0 (no uncertainty regarding the value of $X$), as expected.

*[1]* **(d)** The conditional entropy is:

$$\mathbb{H}(X|Y) = \sum_y p(y)\mathbb{H}(X|Y=y)$$
$$= \frac{1}{4}\left(2\cdot\frac{7}{4} + 2 + 0\right) = 1\tfrac{3}{8}$$

The posterior entropy on $X$ decreases (compared to the prior entropy) when averaged over the possible values of $Y$.

*[1]* **(e)** The mutual information between $X$ and $Y$ is given as:

$$\begin{aligned}
\mathbb{I}(X;Y) = \mathbb{KL}(p(x,y)||p(x)p(y)) &= \sum_{y\in Y}\sum_{x\in X} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \\
&= \Big(p_{XY}(1,1)\log_2\frac{p_{XY}(1,1)}{p_X(1)p_Y(1)} + p_{XY}(2,1)\log_2\frac{p_{XY}(2,1)}{p_X(2)p_Y(1)} + \\
&\quad p_{XY}(3,1)\log_2\frac{p_{XY}(3,1)}{p_X(3)p_Y(1)} + p_{XY}(4,1)\log_2\frac{p_{XY}(4,1)}{p_X(4)p_Y(1)} + \\
&\quad p_{XY}(1,2)\log_2\frac{p_{XY}(1,2)}{p_X(1)p_Y(2)} + p_{XY}(2,2)\log_2\frac{p_{XY}(2,2)}{p_X(2)p_Y(2)} + \\
&\quad p_{XY}(3,2)\log_2\frac{p_{XY}(3,2)}{p_X(3)p_Y(2)} + p_{XY}(4,2)\log_2\frac{p_{XY}(4,2)}{p_X(4)p_Y(2)} + \\
&\quad p_{XY}(1,3)\log_2\frac{p_{XY}(1,3)}{p_X(1)p_Y(3)} + p_{XY}(2,3)\log_2\frac{p_{XY}(2,3)}{p_X(2)p_Y(3)} + \\
&\quad p_{XY}(3,3)\log_2\frac{p_{XY}(3,3)}{p_X(3)p_Y(3)} + p_{XY}(4,3)\log_2\frac{p_{XY}(4,3)}{p_X(4)p_Y(3)} + \\
&\quad p_{XY}(1,4)\log_2\frac{p_{XY}(1,4)}{p_X(1)p_Y(4)} + p_{XY}(2,4)\log_2\frac{p_{XY}(2,4)}{p_X(2)p_Y(4)} + \\
&\quad p_{XY}(3,4)\log_2\frac{p_{XY}(3,4)}{p_X(3)p_Y(4)} + p_{XY}(4,4)\log_2\frac{p_{XY}(4,4)}{p_X(4)p_Y(4)}\Big)
\end{aligned}$$

But this is frankly too tedious. Instead, we use equation 6.51 (PMLI) which specifies the MI in terms of entropies we have already computed:

$$\begin{aligned}
\mathbb{I}(X;Y) &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\
&= \frac{7}{4} + 2 - \frac{27}{8} \\
&= \frac{3}{8}
\end{aligned}$$

## Problem 7 *Eigenvectors (6 pts)*

**PMLI** Exercise 7.2 (page 266). Note: you can check your answer with Python/R/Matlab (include the code you used).

Given a matrix

$$\boldsymbol{A} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$

The eigenvalues of $\boldsymbol{A}$ are obtained by solving hte characteristic equation $\det(\lambda\boldsymbol{I} - \boldsymbol{A}) = 0$:

$$\begin{aligned}
\det\begin{pmatrix} \lambda - 2 & 0 \\ 0 & \lambda - 3 \end{pmatrix} &= 0 \\
(\lambda - 2)(\lambda - 3) - 0 &= 0 \\
\therefore \lambda_1, \lambda_2 &= 2, 3
\end{aligned}$$

The eigenvectors are found by solving:

$$(\lambda_i \boldsymbol{I} - \boldsymbol{A})\boldsymbol{u}_i = \boldsymbol{0}$$

$$\text{For } \lambda_1 = 2: \quad \begin{pmatrix} 2-2 & 0 \\ 0 & 2-3 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

For the above equation to be true, $u_{12}$ must equal 0. However, $u_{11}$ can be any non-zero value. Since any $c\boldsymbol{u}$ can be an eigenvector, we choose the normalized vector of length 1. Thus:

$$\boldsymbol{u}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Similarly, for $\lambda_2 = 3$, $\boldsymbol{u}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

## Problem 8 *Linear system of equations (5 pts)*

Reformulate the two equations:

$$2x_1 + 6x_2 = 8$$
$$5x_1 + x_2 = 0$$

as a system of linear equations, and solve it for $[x_1\ x_2]^T$ using linear algebra.

In matrix form, the system of equations is:

$$\begin{bmatrix} 2 & 6 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 0 \end{bmatrix}$$

We solve for $\boldsymbol{x}$ by multiplying both sides with the inverse of the coefficient matrix (you can do this by hand or use a computer to find this):

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 5 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 8 \\ 0 \end{bmatrix} = \frac{1}{2(1) - 6(5)} \begin{bmatrix} 1 & -6 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 8 \\ 0 \end{bmatrix} = -\frac{1}{28} \begin{bmatrix} 8 \\ -40 \end{bmatrix} = \begin{bmatrix} -\frac{2}{7} \\ \frac{10}{7} \end{bmatrix}$$

## Problem 9 *Subgradients (7 pts)*

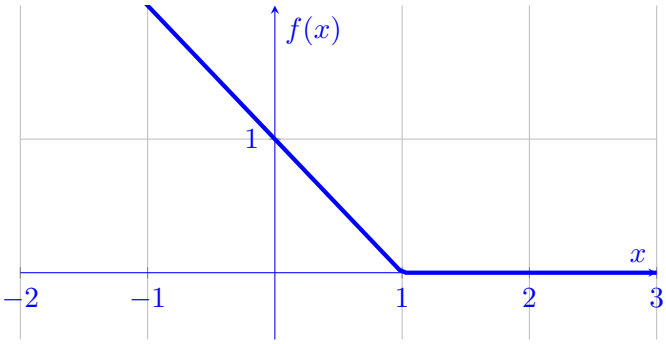
**PMLI** Exercise 8.1 (page 314). [Context: the hinge loss is typically used as a loss function in classifier training, for example in support vector machines.] **Important:** Answer the question along the following steps:

*[2]* **(a)** Sketch/plot the hinge loss function $f(x)$.

The hinge loss function is given by:

$$f(x) = (1 - x)_+ = \max(0, 1 - x)$$

It can be plotted as follows:




**(b)** Write the piecewise subgradient/subderivative $\partial f(x)$ (see equation 8.14 on p. 276 for an example). *[2]*

The hinge loss is given as $f(x) = (1-x)_+$. Its subgradient is:

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 1 \\ [-1, 0] & \text{if } x = 1 \\ \{0\} & \text{if } x > 1 \end{cases}$$

**(c)** Evaluate the subgradient at the specified points in the exercise: $\partial f(0)$, $\partial f(1)$, $\partial f(2)$. *[3]*

$\partial f(0) = -1$
$\partial f(1) \in [-1, 0]$ (Typically, 0 is used)
$\partial f(2) = 0$