- **Section A. Linear Regression** (10 points)

1. (1 point) Linear regression:
   (A) is a form of unsupervised learning
   (B) is a form of supervised learning
   (C) can be used in classification, where the loss function does not differentiate between "easy" and "hard" training samples.
   (D) can be used in classification but is not robust to outliers.

   Your choice(s): BCD

   1 correct answer: 0.33

   2 correct answers: 0.67

   3 correct answers: 1

2. (2 points) You have been hired by a restaurant to model how many people are going to eat there on a given day. The dataset you have consists of pairs (x, y) where x is the number of days since the restaurant opened and y is the number of people who ate there on that day. For example the pair $(0, 61)$ indicates 61 people ate at the restaurant on its opening night. Your task is to use linear regression to predict y as a function of x. Given that you know the number of customers depends primarily on cyclic factors like the season and day of the week, use no more than 2 sentences to describe what might be a suitable choice of features.

   Your answer: (1pt) We can use sine or cosine functions as our choice of feature.

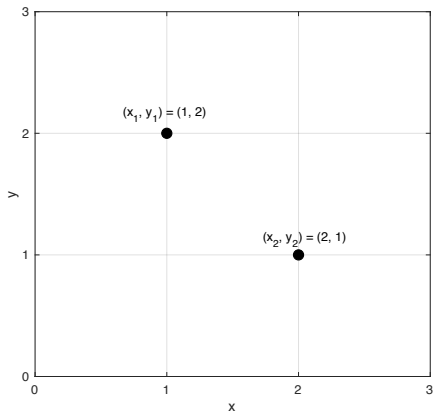
   If students write polynomials, they should receive 0 mark.

3. Assume that a dataset has N training samples $(x_i, y_i), i = 1, ..., N$, where $x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$. We assume the following linear relationship between x and y,

$$y = x + b$$

   In this model, we have only one parameter to optimize, *i.e.,* b.

   (a) (3 points) Using mean squared error, derive the closed-form solution of the parameter b in terms of the training samples x_i and $y_i, i = 1, ..., N$. (Hint: linear regression from the statistics point of view)
   (b) (2 points) Now we have $N = 2$ samples, $(1, 2)$ and $(2, 1)$. Calculate b and the value of the mean squared error in 3a.
   (c) (2 points) Is $y = x + b$ the optimal model for the 2 points in 3b? On the figure below, draw the fitted line that you think is optimal.



Write your answer to 3(a), 3(b), 3(c) on your paper. The fitted line in 3(c) should be drawn on the figure provided in Page 2.

Your answer:

(a)
We first formulate the objective function. That is

$$\begin{aligned}\mathcal{L}(x, y, b) &= \frac{1}{N}\sum_{i=1}^{N}(y_i - (x_i + b))^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}((y_i - x_i) - b))^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2 - \frac{2b}{N}\sum_{i=1}^{N}(y_i - x_i) + b^2\end{aligned}$$

Therefore, we can derive the closed-form solution by calculating the gradient and setting it to zero.

$$\frac{\partial \mathcal{L}}{\partial b} = -\frac{2}{N}\sum_{i=1}^{N}(y_i - x_i) + 2b = 0$$

Therefore,

$$b = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)$$

(b) substitute these two samples into our equations, we have

$$b = \frac{1}{2}((2-1) + (1-2)) = 0$$

and

$$\mathcal{L}(x, y, b) = \frac{1}{2}((2 - (1+0))^2 + (1 - (2+0))^2) = 1$$

(c) No. it should be a straight line go through (1,2) and (2,1) i.e. $y = -x + 3$

- **Section B. Probability** (10 points)

  I have a bag, containing three six-sided dice: a green die g, a red die r, and a blue die b. These dice have different sets of numbers on their faces, as described below,

  $g = \{2, 2, 2, 2, 2, 3\}$
  $r = \{1, 2, 3, 4, 5, 6\}$
  $b = \{0, 1, 1, 1, 8, 11\}$

Figure 1: Two dice (numbers on it are different from our question). "Die" is the singular form of "dice".

(1) (2 point) Whatever number you roll, you get that many dollars. You want to win as much money as possible on average. Which die do you choose, and why?

Your answer:

Since we would get dollars according to the number we roll, we calculate the expected value for each die. Let x be the money we get, then

$$E_g(x) = 2 \times \frac{5}{6} + 3 \times \frac{1}{6} = \frac{13}{6}$$
$$E_r(x) = \sum_{i=1}^{6} i \times \frac{1}{6} = \frac{21}{6} = 3.5$$
$$E_b(x) = 0 \times \frac{1}{6} + 1 \times \frac{3}{6} + 8 \times \frac{1}{6} + 11 \times \frac{1}{6} = \frac{22}{6} = \frac{11}{3}$$

As $E_b(x) > E_r(x) > E_g$ and we want to win as much as possible, we should choose the blue die.

(2) (2 points) What would be a reasonable way to define how random a die is? Justify why you think your definition is a suitable definition of randomness. Which die under your definition is the most random one?

Your answer:

We can use the standard deviation of each die as an indicator. That is

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

where $N = 6$ for all the dice, x_i is number on each face, and μ is the mean calculated in (1). The larger the variance, the more uncertain the die is, so it is more random. Under this definition, we have

$$\sigma_g = \frac{5}{36},\ \sigma_r = \frac{35}{12},\ \sigma_b = \frac{161}{9}$$

Therefore, the blue die is the most random one.

Some equivalent measures (e.g. variance) are also valid.

(3) (2 points) Suppose I choose the blue die, and you chose the green die. We both roll our dice, and whoever has the highest number wins. With what probability do you win this game?

Your answer:

Let x be a random variable represents the number we roll, then we calculate the probability by enumerating all the cases. That is

$$\begin{aligned}
&\mathrm{P}_g(x=2)\mathrm{P}_b(x \in \{0,1\}) + \mathrm{P}_g(x=3)\mathrm{P}_b(x \in \{0,1\}) \\
&= \frac{5}{6} \times \frac{4}{6} + \frac{1}{6} \times \frac{4}{6} \\
&= \frac{2}{3}
\end{aligned}$$

Therefore, the probability that I win is $\frac{2}{3}$

(4)

1. (4 points) Suppose I put the three dice in a bag, shake them up, and draw one out randomly. I roll the die, and you observe the result to be a 1. Suppose furthermore that you are colourblind, and can't tell the colour of the dice apart.[1] With what probability did I select the blue die from the bag? The red die? The green die?

   (Hint: You can shortcut a lot of work once you know the probability of choosing the blue die.)

[1] Otherwise this is a rather trivial question to answer.

Your answer:

let $d \in \{g, r, b\}$ be a random variable represent the color of the chosen die and x be a random variable indicates the number we roll. Since the green die doesn't have 1 on its faces, $P(d = g) = 0$

We calculate the probability by Bayes' Theorem. That is

$$p(d = r|x = 1) = \frac{p(x = 1|d = r)p(d = r)}{p(x = 1)} = \frac{1/6 \times 1/2}{4/12} = 1/4$$
$$p(d = b|x = 1) = 1 - p(d = r|x = 1) = 3/4$$

Therefore, given the result is 1, the probability is $\mathrm{P}(d = g) = 0, \mathrm{P}(d = r) = 1/4, \mathrm{P}(d = b) = 3/4$

- **Section C. Gaussian Mixture Models** (10 points)

1. (1 point) Let n index the data points and k index the mixture components. The parameters r_{nk} in Gaussian Mixture models are
   (A) Used identically to the r_{nk} in K-means clustering.
   (B) Conceptually similar to the r_{nk} in K-means clustering, but can take on more values.
   (C) Completely unrelated to the r_{nk} in K-means clustering.
   (D) Updated iteratively during training.

   Your choice(s):

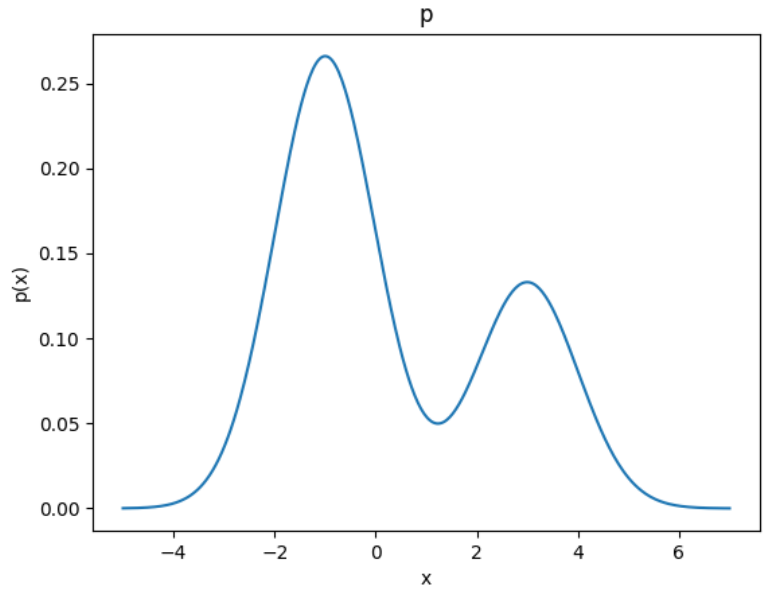
   BD

2. (1 point) Let K denote the number of Gaussian components. How do we determine the parameter K in a GMM?
   (A) It's a hyper-parameter so we use prior knowledge and empirical results
   (B) Gradient descent
   (C) The EM algorithm
   (D) There is a closed form solution that involves inverting a matrix

   Your choice(s):

   A

3. Consider the pictured probability density function p:



We want to approximate p as a gaussian mixture model.

(1) (1 point) What would be a suitable choice of K? Use one sentence to justify your answer.

Your answer: K=2, because there are two peaks.

(2) (1 point) For each $k \in \{1, \ldots, K\}$, suggest a suitable choice of μ_k (approximately).

Your answer: $\mu_1 = -1, \mu_2 = 3$.

(3) (2 points) Suppose we only observe two points $x_1 = -1$, $x_2 = 2$. Their responsibilities r_{nk} are: $r_{11} = 0.9$, $r_{21} = 0.1$. Calculate μ_1 and μ_2.

Your answer: μ_1 and μ_2 are worth 1 pt each. Correctly writing down the formula, but with incorrect calculations will get 1pt.

(4) (2 points) Given that p is a probability density function, what is an advantage of approximating it with a Gaussian mixture rather than linear regression? (no more than 3 sentences)

Your answer: Possible answers:

1) In its nature, linear regression does not predict probabilities.

2) It is hard to find an appropriate feature for linear regression.

3) Every component has a marginal distribution as a Gaussian.

1 correct advantage is worth 1 pt

1 incorrect advantage will lose 1 pt

(5) (2 points) Suppose we observe three points from this distribution, $x_1 = -1.5$, -1.8, and $x_3 = 3$. We use k-means and GMM to cluster these points, respectively, using the same K. Will these two algorithms give us the same means? Explain your answer in 3 sentences.

Your answer:

No.

1) k-means is hard assignment, and GMM is soft assignment.

2) From the three points, x_1 and x_2 are one cluster (component), and x_3 is another cluster (component).

3) For GMM, x_3 will always contribute to the mean of x_1 and x_2, while for k-means, x_3 does not contribute to the mean between x_1 and x_2.

- **Section D. Matrix Decomposition** (10 points)

1. (2 points) For what values of $x \in \mathbb{R}$ is the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 3 \\ 0 & x & -5 \\ 3 & 4 & x \end{bmatrix}$$

invertible? Write your answer on your paper.

Your answer:

A is invertible equivalent to A is full rank, so we derive the condition via Gaussian elimination.

$$\begin{bmatrix} 1 & 0 & 3 \\ 0 & x & -5 \\ 3 & 4 & x \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 0 & 3 \\ 0 & x & -5 \\ 0 & 4 & x-9 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 0 & 3 \\ 0 & x & -5 \\ 0 & 4 & x-9 \end{bmatrix} \rightsquigarrow$$

When $x = 0$, we have

$$\rightsquigarrow \begin{bmatrix} 1 & 0 & 3 \\ 0 & 0 & -5 \\ 0 & 4 & -9 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 1 & 0 & 3 \\ 0 & 4 & -9 \\ 0 & 0 & -5 \end{bmatrix}$$

Which is full rank.

When $x \neq 0$, we have

$$\rightsquigarrow \begin{bmatrix} 1 & 0 & 3 \\ 0 & x & -5 \\ 0 & 0 & x-9+\frac{5\times 4}{x} \end{bmatrix}$$

To make it full rank,

$$x - 9 + \frac{5 \times 4}{x} \neq 0$$
$$x^2 - 9x + 20 \neq 0$$
$$(x-4)(x-5) \neq 0$$

So

$$x \neq 4,\ x \neq 5$$

In conclusion, when $x \in \mathbb{R}\backslash\{4,5\}$, A is invertible

2. Let

$$\mathbf{B} = \begin{bmatrix} 4 & -3 \\ 2 & -3 \end{bmatrix}$$

(1) (2 points) Find all eigenvalues of $\mathbf{B}$.
(2) (2 points) Find a set of eigenvectors of $\mathbf{B}$ that spans $\mathbb{R}^2$.
(3) (2 points) Compute the eigendecomposition of $\mathbf{B}$.
(4) (2 points) Describe how you could use the eigendecompositon of $\mathbf{B}$ to quickly compute $\mathbf{B}^n$ for large integers n. (You don't need to give the closed form expression for $\mathbf{B}^n$.) Write your answer on your paper.

Your answer:

(1) We form the characteristic polynomial and set it to zero. That is

$$p_{\mathbf{B}}(\lambda) = \begin{vmatrix} 4-\lambda & -3 \\ 2 & -3-\lambda \end{vmatrix} = (4-\lambda)(-3-\lambda)+6 = (\lambda-3)(\lambda+2)$$

Therefore, the eigenvalues of $\mathbf{B}$ are $3,\ -2$

(2) we find the corresponding eigenvector of each *eigenvalue* by solve $(A-\lambda\mathbf{I})x = 0$ where $x \in \mathbb{R}^2$

For $\lambda = 3$, we obtain

$$\begin{bmatrix} 4-3 & -3 \\ 2 & -3-3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & -3 \\ 2 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

We solve this homogeneous system and obtain the first eigenvector $p_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$

For $\lambda = -2$, we obtain

$$\begin{bmatrix} 4+2 & -3 \\ 2 & -3+2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 & -3 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

We solve this homogeneous system and obtain the second eigenvector $p_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

(3) The eigenvectors p_1, p_2 form a basis of $\mathbb{R}^2$. Therefore, we can diagonalize B as

$$\mathbf{B} = PDP^{-1}$$

where

$$P = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

$$P^{-1} = \frac{1}{5} \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}$$

and

$$D = P^{-1}\mathbf{B}P = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$$

(4) We derive the computation of $\mathbf{B}^n$ for large integer n via eigendecomposition as below

$$\begin{aligned} \mathbf{B}^n &= (PDP^{-1})^n \\ &= PDP^{-1}PDP^{-1}...PDP^{-1} \\ &= PD^nP^{-1} \end{aligned}$$

Since D is a diagonal matrix, it is easy to compute D^n.

- **Section E. Principal Component Analysis** (10 points)

  You are doing some research that uses PCA to analyze cities. You choose to describe a city using the following features: population, area (km^2), average salary (k AUD/year), temperature (°C). Each column represents a sample, and each row is a feature. From left to right, the columns denote Sydney, Melbourne, Perth and Canberra, respectively.

$$\boldsymbol{X} = \begin{bmatrix} 4600000 & 4800000 & 2000000 & 400000 \\ 12368 & 9990 & 6418 & 814.2 \\ 68.2 & 62.3 & 63.4 & 74 \\ 21.8 & 20.4 & 24.7 & 20.2 \end{bmatrix}$$

  After data standardization, we obtain the eigenvectors and eigenvalues from the data covariance matrix. The standardized matrix $\boldsymbol{A}$, covariance matrix $\boldsymbol{S}$ are shown below.

$$\boldsymbol{A} = \begin{bmatrix} 0.75 & 0.89 & -0.45 & -1.19 \\ 0.99 & 0.52 & -0.19 & -1.31 \\ 0.23 & -0.88 & -0.67 & 1.32 \\ 0.01 & -0.66 & 1.41 & -0.76 \end{bmatrix}$$

$$\boldsymbol{S} = \begin{bmatrix} 0.40 & 0.24 & -0.17 & -0.48 \\ 0.24 & 0.57 & -0.21 & -0.60 \\ -0.17 & -0.21 & 0.70 & -0.29 \\ -0.48 & -0.60 & -0.29 & 1.36 \end{bmatrix}$$

  The eigenvalues and their corresponding eigenvectors (columns) are:

$$\begin{bmatrix} 1.26 \times 10^{-12} & 0.23 & 0.89 & 1.88 \end{bmatrix}$$

$$\boldsymbol{U} = \begin{bmatrix} 0.50 & 0.75 & 0.27 & -0.33 \\ 0.50 & -0.65 & 0.37 & -0.43 \\ 0.50 & -0.06 & -0.86 & -0.08 \\ 0.50 & -0.04 & 0.22 & 0.84 \end{bmatrix}$$

(1) (2 points) Explain the necessity of doing data standardization in this example (no more than 3 sentences).

Your answer:

The data has different scales for different features.

If standardization is not performed, the largest variances might be found at features with the largest scale, which might not be true.

(2) (3 points) By looking at the covariance matrix $\mathbf{S}$, list three observations (each with no more than 2 sentences) you can make.

Your answer:

Population and area are positively correlated.

population and salary are negatively correlated.

Salary and temperature are negatively correlated.

Area and salary are negatively correlated.

Each correct observation is worth 1 point.

An incorrect observation will lose 1 point.

(3) (2 points) If you want to capture 90% of the data variance, which eigenvalue(s) should you choose?

Your answer:

The student should find the largest eigenvalues that contributes to 90% of the sum of all the eigenvalues.

The student will get 2 points if they correctly choose 0.89 and 1.88, which contribute to 92% of the total variance.

If a student knows the captured variance corresponds to value of eigenvalues, he will get 0.5.

If a student incorrectly calculates the percentage of the eigenvalues, he will get 0.5.

(4) (3 points) If we project **A** onto the first principal component (corresponding to the largest eigenvalue), we obtain the following coordinates, -0.68, -1.00, 1.47, 0.22 for Sydney, Melbourne, Perth and Canberra, respectively. Therefore, Perth has the largest coordinate. Describe Perth (regarding its population, area, salary, and temperature) based on it (no more than 4 sentences). (Hint: does Perth have large population among the four cities?)

Your answer:

Among the 4 cities, perth has a relatively small population.

Perth has a relatively small area.

Perth has a relatively low salary.

Perth has a relatively high temperature.

Each correct observation is worth 1 point.

An incorrect observation will lose 1 point.

- **Section F. Classification** (10 points)

1. (1 point) Given N data points $\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, ..., N$, you are using logistic regression for a binary classification problem and can obtain both training accuracy and testing accuracy. Now, you add a new dimension of feature to each data point, such that $\tilde{\boldsymbol{x}}_i = [\boldsymbol{x}_i, u_i] \in \mathbb{R}^{D+1}$. You re-train the classifier until convergence and obtain the new training accuracy and testing accuracy. Select the correct statement(s) below.

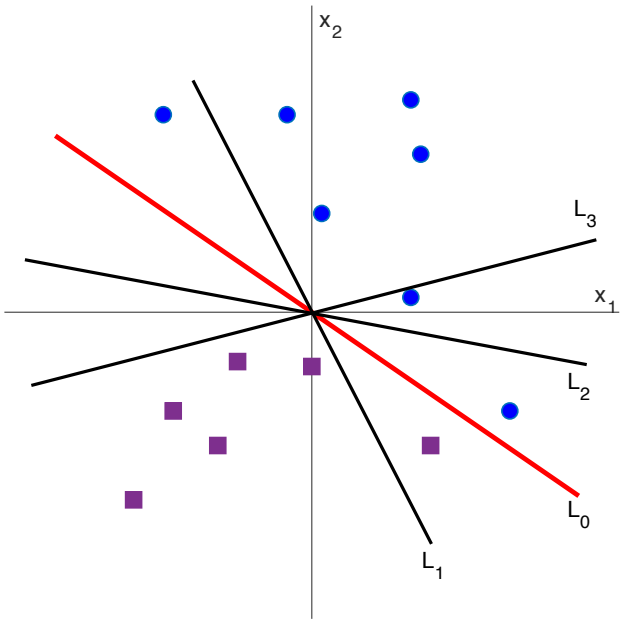
   (A) The new training accuracy is lower than the previous training accuracy.
   (B) The new training accuracy is no lower than the previous training accuracy.
   (C) The new testing accuracy is lower than the previous testing accuracy.
   (D) The new testing accuracy is no lower than the previous testing accuracy.

Your choice(s):
B

2. We are interested in a 2-class classification problem shown below. We'd like to use the logistic regression:

$$h(\boldsymbol{x}; \theta_1, \theta_2) = \mathbb{P}(y = +1 | \boldsymbol{x}, \theta_1, \theta_2) = \frac{1}{1 + \exp\left(-\theta_1 x_1 - \theta_2 x_2\right)}.$$

   We successfully find the decision boundary L_0 (red line).



   (1) (1 point) Can a Perceptron classifier find a suitable decision boundary? Use one sentence to justify your answer.

Your answer: Yes, because the dataset is linearly separable.

   (2) Consider the regularized logistic regression. We maximize the following function.

$$\sum_{i=1}^{N} \log \mathbb{P}(y_i | \boldsymbol{x_i}, \theta_1, \theta_2) - \frac{\lambda}{2}\theta_1^2.$$

   where $\lambda \geq 0$ is the weighting parameter. We re-train the classifier using this equation. Answer the following questions regarding the new decision boundary.

(a) (2 points) Will the decision boundary be L_1? Provide a short justification (no more than 3 sentences).

Your answer: No.

The decision boundary is $x_2 = -\frac{\theta_1}{\theta_2} x_1$

When the magnitude of θ_1 is constained to be minimized, the absolute value of the slope $-\frac{\theta_1}{\theta_2}$ becomes smaller, so the decision boundary becomes flatter.

(b) (2 points) Will the decision boundary be L_2? Provide a short justification (no more than 3 sentences).

Your answer: L_2 is possible.

When θ_1 is constrained to be small, the new decision boundary should be flatter than L_0.

L_2 is flatter than L_1.

Compared with L_3 which has a similar slope, the training error of L_2 is smaller, so L_2 is possible.

Note: if the first point is answered in question (a), then student will automatically get 0.5 point.

Note: if the third point is answered in (c), the student will automatically get 0.5 point.

(c) (2 points) Will the decision boundary be L_3? Provide a short justification (no more than 3 sentences).

Your answer: L_3 is flatter than L_0, and has a similar slope magnitude with L_2.

(1 pt) Comparing L_3 and L_2, we find L_3 has a larger training error (two blue points are misclassifed) than L_2 (1 blue point is misclassified).

Therefore, L_3 will at least converge to L_2 where the overall loss is smaller.

(3) (2 points) If we increase the value of λ starting from $\lambda = 0$, how will the average log-probability on the *test set* change? Use no more than 5 sentences to justify your answer.

Your answer:

When λ is relatively small, regularization is effective, so usually the test accuracy (log-probability on the test set) will become higher.

When λ beomes relatively large, regularization effect is unnecessarily large, so usually the test accuracy (log-probability on the test set) will become lower.

——— End of the paper ———