Problem 1: Background [25pt]

1. **Define the terms in a couple of sentences with your own words (10pt):**

- Generalization:

A model's ability to adapt and give the right result even on the data that it hasn't been trained on, is called Generalization. If a model isn't able to generalize well, it might give accurate results on training set, but it fails to do well on unseen test set.

- Overfitting:

If we try to have an overly complex model in order to get a high accuracy on the training data, the model will end up learning the training data too well. Then when the same model is used on the testing dataset, it will not be able to give a good result because it has adjusted itself to the training set completely. This problem is called overfitting. As the word indicates, overly fitting a model to the training dataset.

- Underfitting:

If a model is too simple, it won't even learn the training dataset well. It will give a very low accuracy on both training and testing datasets. This problem is called under fitting.

- Regularization:

Regularization is the process of ensuring that a model does not become too complex and starts memorizing the data(avoid overfitting). This is done by introducing a regularization term in the model.

- No free lunch theorem:

No free lunch theorem explains that there is no one particular model that will fit all types of data. If a model is performing very well on a certain type of data, it is also possible that the same model will perform very poorly on another type of data. Therefore, what model to use, completely depends on the data.

- Occam's razor:

This theory suggests that if we have to choose between simple and complex models, always choose the simple ones. The more complex a model gets, the less likely it is to generalize well.

- Independent and identically distributed data points:

I.i.d is one of the assumptions that we make about data when we try to use Machine Learning on it. It means that the samples collected are Independent of each other and they are coming from the same distribution. Independent of each other means that one sample does not in any way affect another. For ex, if we are collecting data about students' scores, the score of one student does not affect another's. 'Identically distributed' in this example would mean that all the data samples are related to scores only.

- Cross-validation

If we train a model with the entire dataset, we won't know how exactly the model will perform on unseen data. To solve this problem, we save a part of the dataset as Test/Validation set and use it to evaluate the model by doing an error estimation. This process is called cross-validation.

- Degrees of freedom

This refers to the number of parameters in a model that are allowed to vary. For example, in Linear regression of degree 2 ($y=W0+w1X+w2X^2$), W0, W1 and W2 are the variables. So the degree of freedom is 3.

2. **Assume that you observe two different coins being tossed as follows (5pt):**

   Coin 1 =H,H,H,H,T,T,H,H,H,H,T,T,H,H,H,H,T

   Coin 2 =H,H,T,T,T,T,H,H,T,T,T,T,H,H,T,T,T

   Assume the coin tosses are i.i.d. random variables. Each coin will be tossed one more time and you will be given $100 for each correct guess. What is your guess for Coin 1 and Coin 2 's next toss and why?

   NOTE: There is no one correct solution for this problem. You are free to interpret it in any way you wish. Please make sure to list all the assumptions that you make in the context of this question.

   Since the coins are I.i.d, the toss of one coin does not affect the other. Also, each toss is independent and unaffected by the previous toss. Thus every toss of each coin has 0.5 probability of landing on head and 0.5 probability of landing on tail.

   I will guess that coin 1 will give Tail and coin 2 will also give tail. The reason is that there seems to be a pattern of 4 H followed by 2 T's in the distribution of tosses of coin1 and the pattern of H's followed by 4 T's in the distribution of tosses of coin 2. I would have a 50% chance of winning $100 for each of these guesses.

3.

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \left[y(x_n, w) - t_n \right]^2$$

$$y(x, w) = w_0(1) + w_1(x) + w_2(x)^2 + \dots + w_m x^m = \sum_{j=0}^{m} w_j x^j$$

$$= \begin{bmatrix} w_0 & w_1 & w_2 & \cdots & w_m \end{bmatrix} \begin{bmatrix} x_1 \\ x^2 \\ x^3 \\ x^{\sim} \\ x \\ \vdots \\ x^n \end{bmatrix}$$

$$= w^T x \qquad \text{where,} \quad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}$$

$$\therefore E(w) = \frac{1}{2} \sum_{n=1}^{N} \left[w^T x_n - t_n \right]^2$$

$$= \frac{1}{2} \sum_{n=1}^{N} \left[x_n^T w - t_n \right]^2 \qquad \because A^T B = B^T A$$

$$= \frac{1}{2} \left(\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & & \ddots & \vdots \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Nm} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix} - \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} \right)^2$$

$$= \frac{1}{2} (Xw - t)^2 \qquad \text{where,} \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ \vdots & & \ddots & \vdots \\ \vdots & & & \\ x_{N1} & & \cdots & x_{Nm} \end{bmatrix} \quad t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

$$= \frac{1}{2} (X.w - t)^T (Xw - t)$$

$$= \frac{1}{2} ((Xw)^T - t^T)(Xw - t)$$

$$= \frac{1}{2} (w^T X^T - t^T)(Xw - t)$$

$$= \frac{1}{2}\left(w^T X^T (Xw - t) - t^T (Xw - t)\right)$$

$$= \frac{1}{2} w^T X^T X w - \frac{1}{2} (Xw)^T t - \frac{1}{2} t^T (Xw) + \frac{1}{2} t^T t$$

$$= \frac{1}{2} w^T X^T X w - w^T X^T t + \frac{1}{2} t^T t$$

$$\frac{\partial}{\partial w}\left(E(w)\right) = 0$$

$$\Rightarrow \quad X^T X w - X^T t = 0$$

$$= \quad X^T X w = X^T t$$

$$\Rightarrow \quad \boxed{w_* = (X^T X)^{-1} X^T t}$$

$$y = X w_* \quad \Rightarrow \quad \boxed{y = X (X^T X)^{-1} X^T t}$$