**Questions:**

1. **Can you use logistic regression for classification between more than two classes?**
2. **3If you observe that the cost function decreases rapidly before increasing or stagnating at a specific high value, what could you infer?**
3. **How many binary classifiers would you need to implement one-vs-one for four classes? How does it work?**
4. **How do you implement multinomial logistic regression?**
5. **What alternative could you suggest using a for loop (which is time-consuming) when using Gradient Descent to find the optimum parameters for logistic regression?**
6. **Are there alternatives to find optimum parameters for logistic regression besides using Gradient Descent?**
7. **Why is the Wald Test useful in logistic regression but not in linear regression?**
8. **What are odds? Why is it used in logistic regression?**
9. **Given fair die, what are the odds of occurrence of odd numbers?**
10. **It is common practice that when the number of features or independent variables is larger in comparison to the training set, it is common to use logistic regression or support vector machine (SVM) with a linear kernel. What is the reason behind this?**

**Answers:**

1. **Can you use logistic regression for classification between more than two classes?**

Yes, it is possible to use logistic regression for classification between more than two classes, and it is called multinomial logistic regression. However, this is not possible to implement without modifications to the vanilla logistic regression model.

1. **If you observe that the cost function decreases rapidly before increasing or stagnating at a specific high value, what could you infer?**

A trend pattern of the cost curve exhibiting a rapid decrease before then increasing or stagnating at a specific high value indicates that the learning rate is too high. The gradient descent is bouncing around the global minimum but missing it owing to the larger than necessary step size.

1. **How many binary classifiers would you need to implement one-vs-one for four classes? How does it work?**

To implement one-vs-one for four classes, you will require six binary classifiers. This is because you will need to compare each class with each other class. In general, the formula for calculating the number of binary classifiers b is given as b=(no. of classes \* (no. of classes -1))/ 2.

Suppose we have four different categories into which we need to classify the weather for a particular day: Sun, Rain, Snow, Overcast. Then to implement the one-vs-one approach, we need to make the following comparisons:

    Binary Classification Problem 1: Sun vs. Rain

    Binary Classification Problem 2: Sun vs. Snow

    Binary Classification Problem 3: Sun vs. Overcast

    Binary Classification Problem 4: Rain vs. Snow

    Binary Classification Problem 5: Rain vs. Overcast

    Binary Classification Problem 6: Snow vs. Overcast

1. **How do you implement multinomial logistic regression?**

The multinomial logistic classifier can be implemented using a generalization of the sigmoid, called the softmax function. The softmax represents each class with a value in the range (0,1), with all the values summing to 1. Alternatively, you could use the one-vs-all or one-vs-one approach using multiple simple binary classifiers.

1. **What alternative could you suggest using a for loop (which is time-consuming) when using Gradient Descent to find the optimum parameters for logistic regression?**

One commonly used efficient alternative to using for loop is vectorization, i.e., representing the parameter values to be optimized in a vector. By using this approach, all the vectors can be updated instead of iterating over them in a for loop.

1. **7Are there alternatives to find optimum parameters for logistic regression besides using Gradient Descent?**

Yes, [Gradient Descent](https://www.projectpro.io/article/the-a-z-guide-to-gradient-descent-algorithm-and-its-variants/434) is merely one of the many available optimization algorithms. Other advanced optimization algorithms can often help arrive at the optimum parameters faster and help with scaling for significant machine learning problems. A few such algorithms are Conjugate Gradient, BFGS, and L-BFGS algorithms.

1. **Why is the Wald Test useful in logistic regression but not in linear regression?**

The Wald test, also known as the Wald Chi-Squared Test, is a method to find whether the independent variables in a model are of significance. The significance of variables is decided by whether they contribute to the predictions or not. The variables that add no value to the model can therefore be deleted without risking severe adverse effects to the model. The Wald test is unnecessary in linear regression because it is easy to compare a more complicated model to a simpler model to check the influence of the added independent variables. After all, we can use the R2 value to make this comparison. However, this is not possible with logistic regression as we use Maximum Likelihood Estimate, which uses the previously mentioned method infeasible. The Wald test can be used for many different models, including those with binary variables or continuous variables, and has the added advantage that it only requires estimating one model.

1. **What are odds? Why is it used in logistic regression?**

Odds are the ratio of the probability of success to the probability of failure. The odds serve to provide the constant effect a particular predictor or independent variable has on the output prediction. Expressing the effect of a predictor on the likelihood of the target having a particular value through probability does not describe this constant effect. In linear regression models, we often want to measure the unique effect of each independent variable on the output for which the odds are very useful.

1. **Given fair die, what are the odds of occurrence of odd numbers?**

Theodds of occurrence of odd numbers is 1.

There are three odd and three even numbers in a fair die, and therefore, the probability of occurrence of odd numbers is 3/6 or 0.5. Similarly, the odds of occurrence of numbers that are not odd is 0.5. Since odds is the ratio of the probability of success and that of failure,

Odds = 0.5/0.5=1.

1. **It is common practice that when the number of features or independent variables is larger in comparison to the training set, it is common to use logistic regression or support vector machine (SVM) with a linear kernel. What is the reason behind this?**

It is common to use logistic regression or SVM with a linear kernel because when there are many features with a limited number of training examples, a linear function should be able to perform reasonably well. Besides, there is not enough training data to allow for the training of more complex functions.