Q1

When you duplicate feature n into feature $(n+1)$ and retrain a logistic regression model, the likely relationship between the weights $w_{\text{new}0}$, $w_{\text{new}1}$, $w_{\text{new}n}$, and $w_{\text{new}n+1}$ can be explained based on how logistic regression works.

In logistic regression, the weights (w) are learned to minimize a loss function, typically the logistic loss or cross-entropy loss. Each weight is associated with a feature, and it represents the contribution of that feature to the prediction. When you duplicate feature n into $(n+1)$, the new feature $(n+1)$ essentially contains the same information as feature n.

In this scenario, considering the duplicate features, you might observe the following relationships:

1. **Weights for the Original Feature (n):**
   - $w_{\text{new}n}$ is likely to be close to w_n.
   - $w_{\text{new}n+1}$ is also likely to be close to w_n because it essentially contains the same information.
2. **Weights for the Duplicate Feature ($(n+1)$):**
   - $w_{\text{new}n}$ is likely to be close to w_n because it's still representing the original feature.
   - $w_{\text{new}n+1}$ is likely to be close to w_n because it's a duplicate of feature n.
3. **Bias Term ($w_{\text{new}0}$):**
   - The bias term ($w_{\text{new}0}$) may change during training based on the new data, but it's not directly affected by the duplication of features.

In summary, after retraining the model with the duplicated feature, you can expect the weights associated with the duplicated feature ($w_{\text{new}n}$ and $w_{\text{new}n+1}$) to be close to the original weight (w_n), as they essentially represent the same information.

Q2

We need to compare the confidence intervals for the click through rates of each template. A confidence interval is a range of values that is likely to contain the true population parameter with a certain level of confidence. For example, a 95% confidence interval means that we are 95% confident that the true population parameter is within the interval.

To calculate the confidence interval for a proportion, such as the click through rate, we can use the following formula:

$$p \pm z_{\alpha/2}\text{sqrt}(p(1-p)/n)$$

where p is the sample proportion, $z_{\alpha/2}$ is the critical value from the standard normal distribution for the chosen significance level α, and n is the sample size.

In this case, we have α=0.05, which means $z_{\alpha/2}$=1.96. The sample size is n=1000 for each template. The sample proportions are the click through rates for each template. We can plug these values into the formula and get the following confidence intervals:

- Template A: 0.10±1.960.10(1−0.10)1000=(0.087,0.113)0.10±1.9610000.10(1−0.10)=(0.087,0.113)
- Template B: 0.07±1.960.07(1−0.07)1000=(0.058,0.082)0.07±1.9610000.07(1−0.07)=(0.058,0.082)
- Template C: 0.085±1.960.085(1−0.085)1000=(0.073,0.097)0.085±1.9610000.085(1−0.085)=(0.073,0.097)
- Template D: 0.12±1.960.12(1−0.12)1000=(0.107,0.133)0.12±1.9610000.12(1−0.12)=(0.107,0.133)
- Template E: 0.14±1.960.14(1−0.14)1000=(0.127,0.153)0.14±1.9610000.14(1−0.14)=(0.127,0.153)

To compare the templates, we can look at the overlap of the confidence intervals. If two intervals do not overlap, it means that there is a significant difference between the proportions with 95% confidence. If two intervals overlap, it means that there is not enough evidence to conclude that the proportions are different with 95% confidence.

Based on this, we can see that:

- Template E is better than template A with over 95% confidence, because the interval for E is entirely above the interval for A.
- Template B is worse than template A with over 95% confidence, because the interval for B is entirely below the interval for A.
- Template C is not significantly different from template A with 95% confidence, because the interval for C overlaps with the interval for A.
- Template D is better than template A with over 95% confidence, because the interval for D is mostly above the interval for A, except for a very small overlap.

Therefore, the correct answer is option 2. E is better than A with over 95% confidence, B is worse than A with over 95% confidence. You need to run the test for longer to tell where C and D compare to A with 95% confidence.

Q3

In logistic regression, the key computational steps during each iteration are:

→Compute the hypothesis function:

This involves calculating the dot product of the feature vector and the parameter vector.

→Compute the gradient:

This involves calculating the gradient of the logistic loss function.

→Update the parameters:

This involves updating the parameter vector using the computed gradient.

For sparse data, where the average number of non-zero entries in each training example is k, and k << n, certain optimizations can be applied to take advantage of the sparsity of the data. Many modern machine learning packages use sparse matrix representations and algorithms optimized for sparse data.

In general, the time complexity per iteration for logistic regression with sparse data is often approximately O(m * k), where m is the number of training examples and k is the average number of non-zero entries in each example.

This is the cost for one iteration of gradient descent. The total computational cost will also depend on the number of iterations needed for the algorithm to converge.

Q5

1. Maximum Likelihood Estimate (MLE):

   - Estimate: k/n

   - Explanation: MLE is the proportion of observed heads in the total number of coin tosses.

2. Bayesian Estimate:

   - Estimate: The expected value of the posterior distribution of p, given a continuous uniform prior on p from 0 to 1.

   - Explanation: Assume a continuous uniform distribution as a prior for p in the range [0, 1]. The posterior distribution, given the observations, is a Beta distribution with parameters k+1 and n-k+1. The Bayesian estimate for p is the expected value of this Beta distribution, which is k+1/n+2.

3. Maximum a Posteriori (MAP) Estimate:

- Estimate: The mode of the posterior distribution of p, given a continuous uniform prior on p from 0 to 1.

- Explanation: Similar to the Bayesian estimate, the prior is a continuous uniform distribution on p from 0 to 1. The posterior distribution is a Beta distribution with parameters k+1 and n-k+1. The MAP estimate is the mode of this Beta distribution, which is k/n unless k = 0 or n = k, in which case the mode is 0 or 1, respectively.

→To summarize:

- MLE: k/n

- Bayesian Estimate: k+1/n+2

- MAP Estimate: k/n (unless (k = 0 or n = k)