Q1) In logistic regression, when you have two identical features, the effect of multicollinearity comes into play. Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related.

In your case, where feature (n+1) is a duplicate of feature N, the weights learned by the model for these two features will likely be divided across them. The exact distribution would depend on the specific optimization process, but the sum of weights new_n and w_new_n+1 would likely be approximately equal to the original weight w_n in the model without the duplicate feature. This is because the duplicate feature doesn't provide any new information to the model, so the total weight assigned to that information should remain roughly the same.

As for w_new_0, w_new_1, … , w_new_n-1, they would likely be very similar to the corresponding weights in the original model, since the non-duplicated features and their relationship to the output have not changed. The slight differences might be due to the optimization process and how the additional collinearity affects it.

2) To determine which of the statements is true, we need to perform a statistical test such as a z-test, which is commonly used for comparing proportions. However, to perform a full z-test, we would need to know the standard deviation of the click through rates, which we don't have in this situation.

In general, we can't make definitive statements about statistical significance without actually calculating the p-values or confidence intervals. Nonetheless, we can make some educated guesses based on the distance between the observed proportions:

1 This could be true, as we don't have evidence to reject this. A more detailed statistical test is needed to determine if we have enough data to make a conclusion with 95% confidence.

2 Template E (14% CTR) does seem to perform better than Template A (10% CTR), and Template B (7% CTR) does seem to perform worse. However, without performing a statistical test, we cannot confidently make a statement about the degree of confidence.

3 Again, while Templates D and E appear to perform better than A, and B and C appear to perform worse, we cannot make definitive conclusions regarding the confidence levels without a statistical test.

So, based on this, none of the statements can be definitively confirmed as true without performing a proper statistical test. However, given the data, the most plausible statement seems to be option 2, but it would still require a more detailed analysis to determine the statistical significance level.

Q3) In most well-written packages for logistic regression, the computational cost for each iteration of gradient descent is proportional to the number of non-zero entries in the feature matrix. This is because these packages take advantage of the sparsity of the matrix and only perform computations on the non-zero elements.

So, if you have m training examples and each example has on average k non-zero features, the computational cost for each gradient descent iteration would be approximately O(m * k).

This is much more efficient than O(m*n), which would be the computational cost without taking advantage of the sparsity of the matrix. Note that O(m * k) is a simplification and the actual cost could be higher due to overhead and other factors, but this gives a rough idea of the computational complexity.

4) 1. This method might help to improve the classifier's performance in the "uncertain" areas where the V1 classifier's output is closest to the decision boundary. But it may not help in improving the overall accuracy as it does not necessarily cover all types of errors from the V1 classifier.

1 This method helps to get a more representative sample of the data from the 1000 news sources. It is likely to improve the overall generalization of the classifier. However, it might not target specific weaknesses of the V1 classifier.

2 This method targets the areas where the V1 classifier is most confident but wrong. It is likely to improve the classifier's performance on these types of examples. But like method 1, it does not necessarily cover all types of errors or represent the full diversity of the 1000 news sources.

In terms of ranking, it's hard to definitively rank the methods without more information about the specific distributions and characteristics of the news stories. However, based on the general principles of machine learning and the information given, we might speculate that method 2 could potentially provide the greatest improvement in overall accuracy, because it provides the most diverse and representative sample of the data.

Method 3 might be the next best, as it targets the examples where the V1 classifier is most confident but wrong. This could help to correct systematic biases or errors in the V1 classifier.

Method 1 might be the least effective, since it focuses on the examples that are closest to the decision boundary, which may not represent the full range of examples where the classifier makes errors.

Again, these are speculative rankings and the actual performance would depend on various factors such as the distribution and characteristics of the news stories, the specific weaknesses of the V1 classifier, and the capacity of the model.

5) 1. Maximum Likelihood Estimate (MLE): The MLE estimate of p is simply the proportion of heads that you observed, which is k/n.

1 Bayesian Estimate: Here, you assume a uniform prior, which means you initially assume that all values of p from 0 to 1 are equally likely. After observing the data, you update this prior to get the posterior distribution. For a uniform prior and binomial likelihood, the posterior distribution is a Beta distribution with parameters k+1 and n-k+1. The expected value of this distribution is (k+1) / (n+2), so this is the Bayesian estimate of p.

2 Maximum a posteriori (MAP) estimate: The MAP estimate is the mode of the posterior distribution. For a Beta distribution with parameters k+1 and n-k+1, the mode is (k+1-1) / (n+2-2), which simplifies to k / n, the same as the MLE estimate. Note that this applies for $k > 0$ and $n > k$. If $k = 0$, the MAP estimate is 0, and if $n = k$, the MAP estimate is 1.