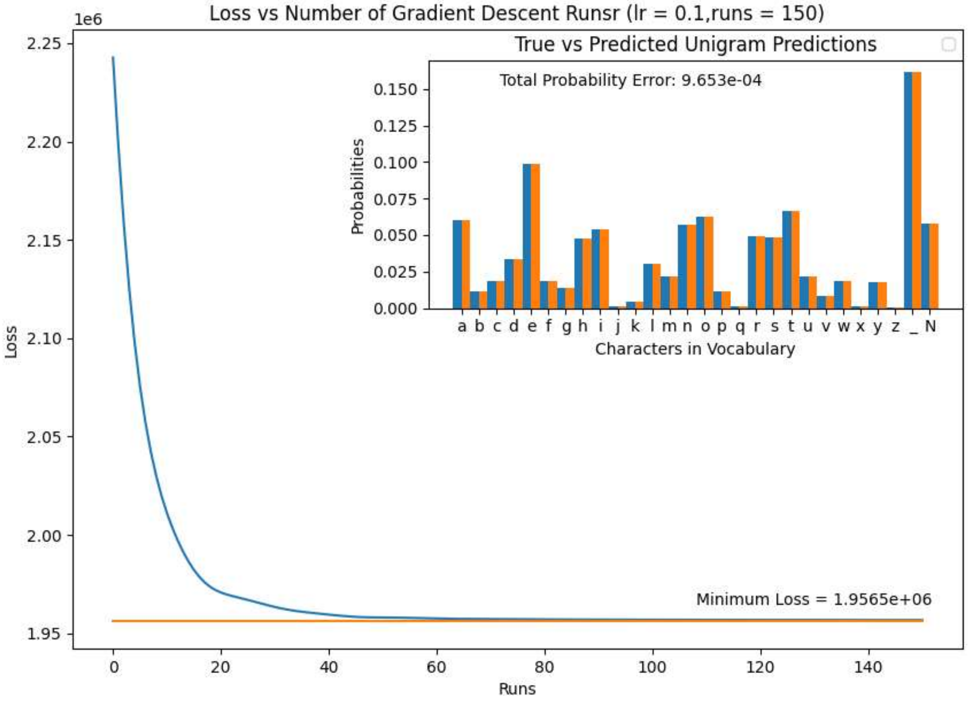
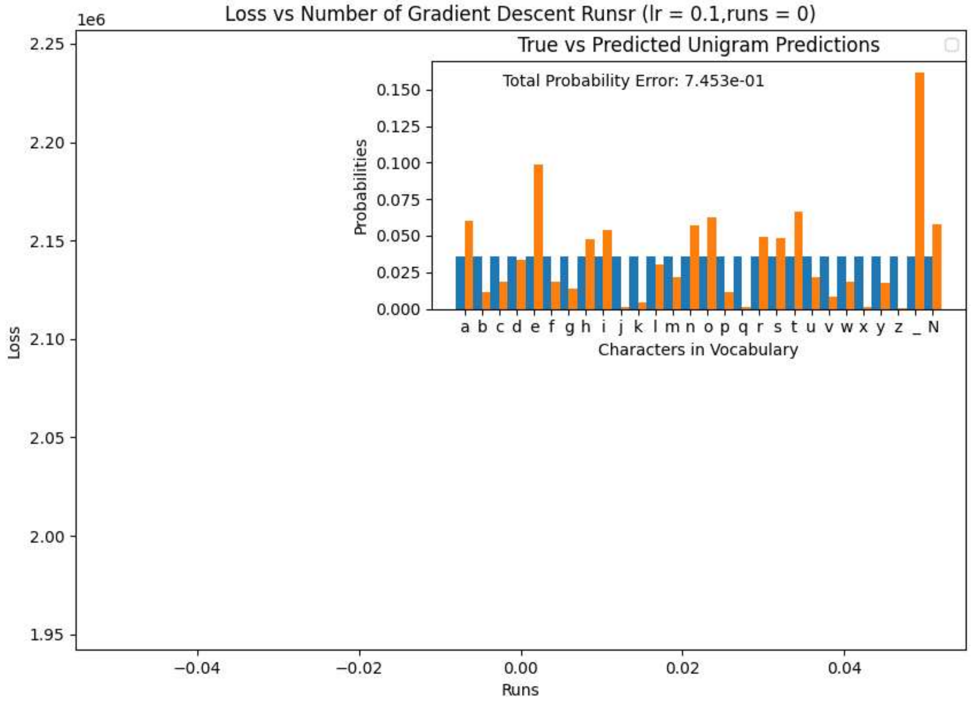
# Gradient Descent Homework

For this assingment we were given an unigram language model that is trained on the Jane Austen corpus, "Sense and Sensibility". We were asked to find an appropriate learning rate and number of iterations for gradient descent. Below I show a plot for the loss as a function of runs with the associated predicted vocabulary probabilities against the true unigram probabilities. The minimum theoretical loss is the computed loss for the true probabilities of the unigrams in the training data. We see convergence towards this value since, if our model is operating perfectly, we should accurately assign all vocabulary characters to their true values with any OOV characters captured in the 'None' token probability.



We can also see that the sum of the differences between true probabilities and predicted probabiliites in this learning rate/number of runs example is very low. This indicates that after 150 runs we have a predicted probabilities very close to the true unigram values.

We can show what this model's initialized probabilities are by plotting the model at 0 iterations:



This shows the initialized p values which are the sigmoid logit of 1/V which approximately 0.03. So we can see that our gradient descent is taking the appropriately initialized predictions, initialized as a uniform distribution, and accurately shifting them towards the desired unigram probabilities.