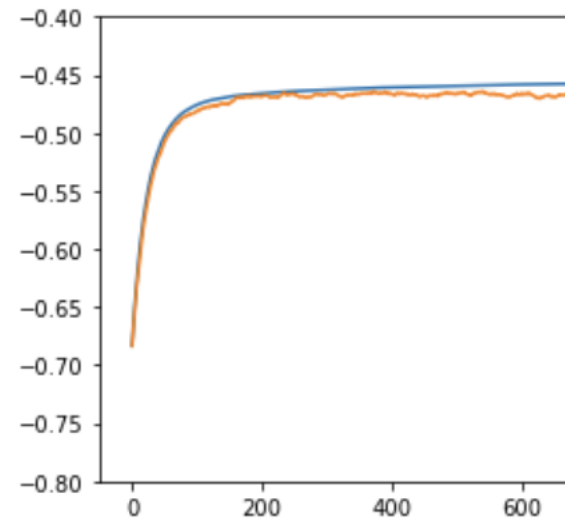
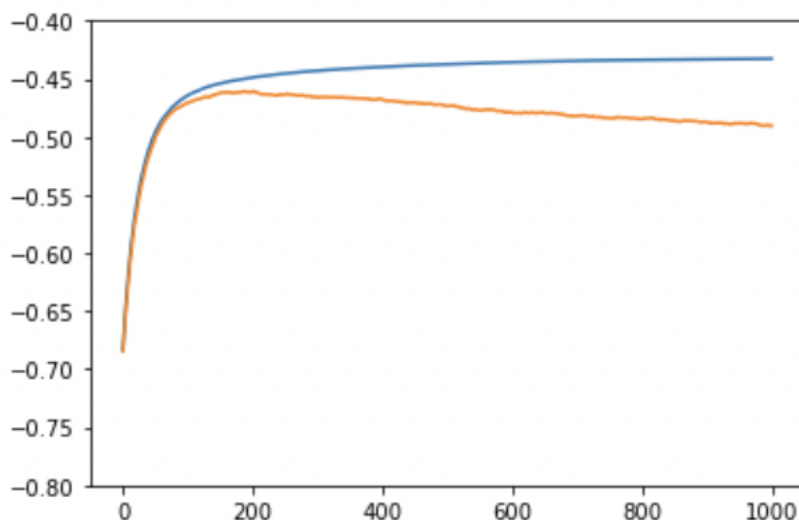


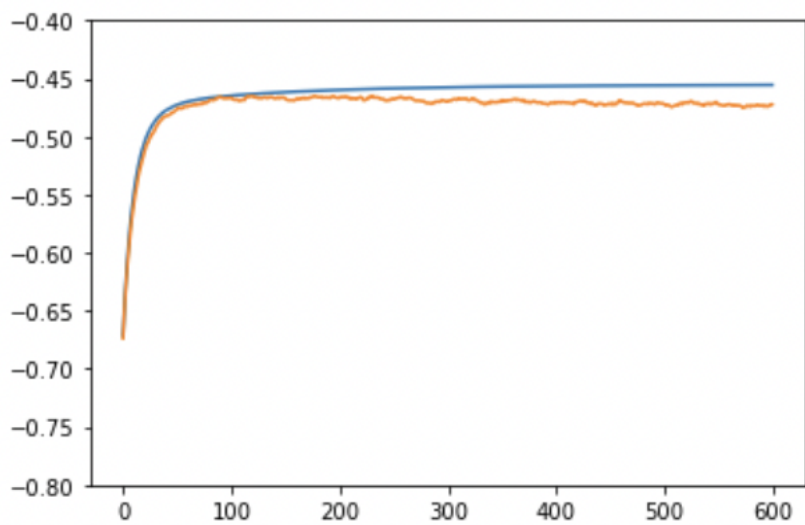
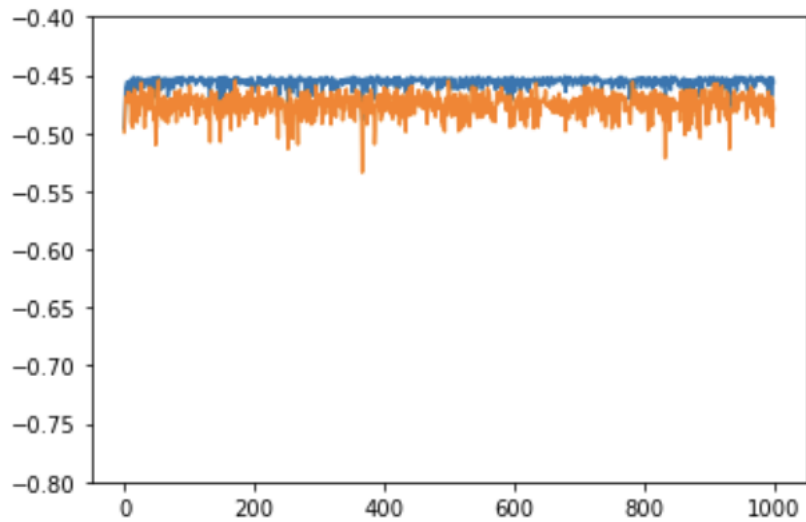
Rachel Springer (rs4127)  
October 8, 2023

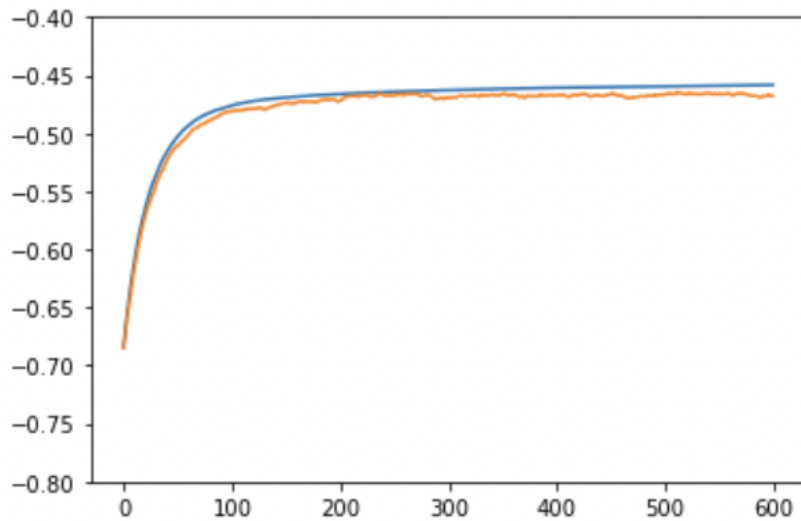
I implemented Bayesian logistic regression on a dataset of Starbucks union elections from August 2021- August 2023. I got the data on their locations, dates, and results from the [unionelections.org](https://unionelections.org) website, and modified the dataset to include attributes that I thought could be relevant or interesting to study. The inputs included: whether or not the union election took place in a "Right to Work" state, the median household income of the town where the election took place in the 2020 census, the minimum wage of the state, the Gini coefficient for the state (measuring income inequality), the percentage union membership in the state, the percentage of Democratic votes from the state in the 2020 presidential election, the margin shift from Republican to Democrat in the 2020 presidential election, the percent of immigrants in the state, whether the ballot was the initial vote or revised vote, the percentage of eligible workers that voted in the election, the total number of eligible voters in the election, the relative date filed (measured from the first date a union election was filed for), the relative tally date (also measured from the first file date), the number of days between filing and voting, and the number of elections that had taken place in that state so far. I found most of this information from the census, though most of the data was limited by the fact that it was much harder to get local data on the town, and so generalized state demographic were used despite clear variances in the social context of different locations within each state. I scaled the features to a mean of 0 and variance of 1 in order to make the results of the regression more interpretable and to improve the results of stochastic gradient descent. This resulted in 451 total observations. The response variable measured the success of the election, or whether the number of votes to join the union was greater than the number of votes to not join the union. The pre-scaled data is attached in the files here:)

I chose a large prior variance for the Gaussian prior within the model because I didn't know the effects of the parameters well. Small variances promoted more overfitting to the test data, whereas larger variances generalized better but performed worse on the train data. Below are graphs with a prior of  $10^{-2}$  vs  $10^2$ .

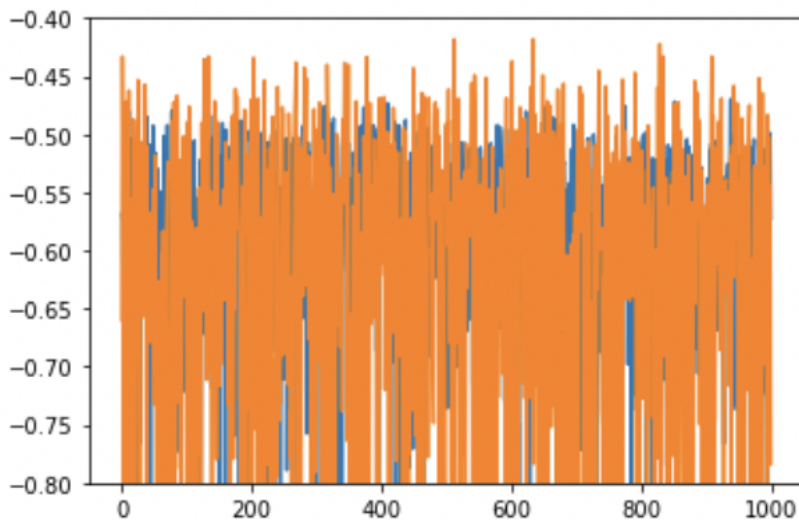


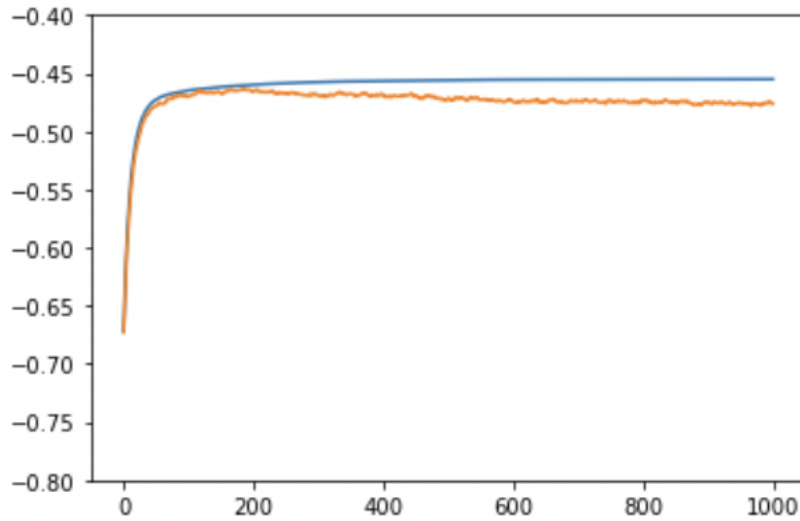
I adjusted batch size, but since my dataset was small used a small batch size for the stochastic gradient descent. Very small batches were unstable (see below, batch sizes of 6, 48, and 128) but large batches took longer to improve. I ultimately chose a batch size of 48 for this reason



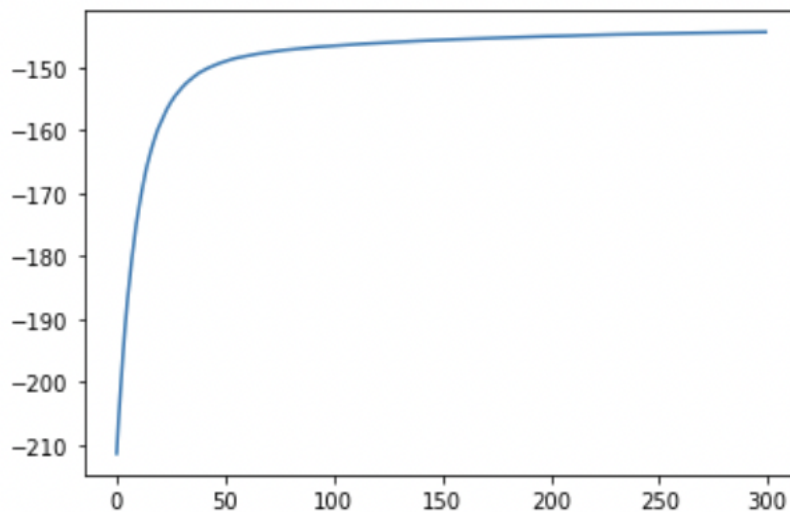


Once I had set these parameters, I decreased the learning rate until the objective function didn't "jump around" as much. Since I was using a smaller batch size, this made my learning rate smaller as well in order to have a more stable system. I found it was smooth enough at 0.0001 (compared to 0.01)





I used a number of iterations, under these parameters, where the test cross entropy loss seemed to peak to prevent overfitting. On this graph, that meant setting the number of iterations to 300 (see above). Using that information, we can plot to log likelihood over the iterations:



The resultant beta for the data was [0.27227037, 0.00954723, -0.23437002, 0.54283922, 0.37180527, 0.58296698, 0.48565024, -0.57440793, 0.1693631, -0.18562403, -0.10718817, 0.08338559, 0.00400394, -0.35841185, -0.14250866] with a bias term of 1.49145598 corresponding to ['RTW', 'MedianIncome', 'MinWage', 'InequalityIndex', 'UnionMembership', 'DemVotes', 'MarginShift', 'IsInitial', 'PercentVoters', 'NumEligibleVoters', 'RelativeDateFiled', 'RelativeTallyDate', 'FileToTallyDays', 'NumElectionsInState'].

This implies the most strong positive correlations between living in a state with more inequality and success in a union election, being in a more Democratic-voting state, being in a state that shifted towards voting more for Democrats than Republicans in the 2020 presidential election. This could be explained by the idea that more exposure to income inequality makes a person more likely to

want to join a union or understand class concepts, and the association between the political left and unionism. On the other hand, states with more immigrants, and elections where it took more time between filing for an election were negatively correlated with success. There is some theorizing that immigrant communities are less likely to join unions for fear of repercussions, which may be reflected here, and an increased time between filing and voting may be related to a more contested election, explaining the negative correlation. Interestingly, the median income of the town and the order of the election (an attempt to measure "momentum") did not seem to have a strong correlation to the outcome of the election.