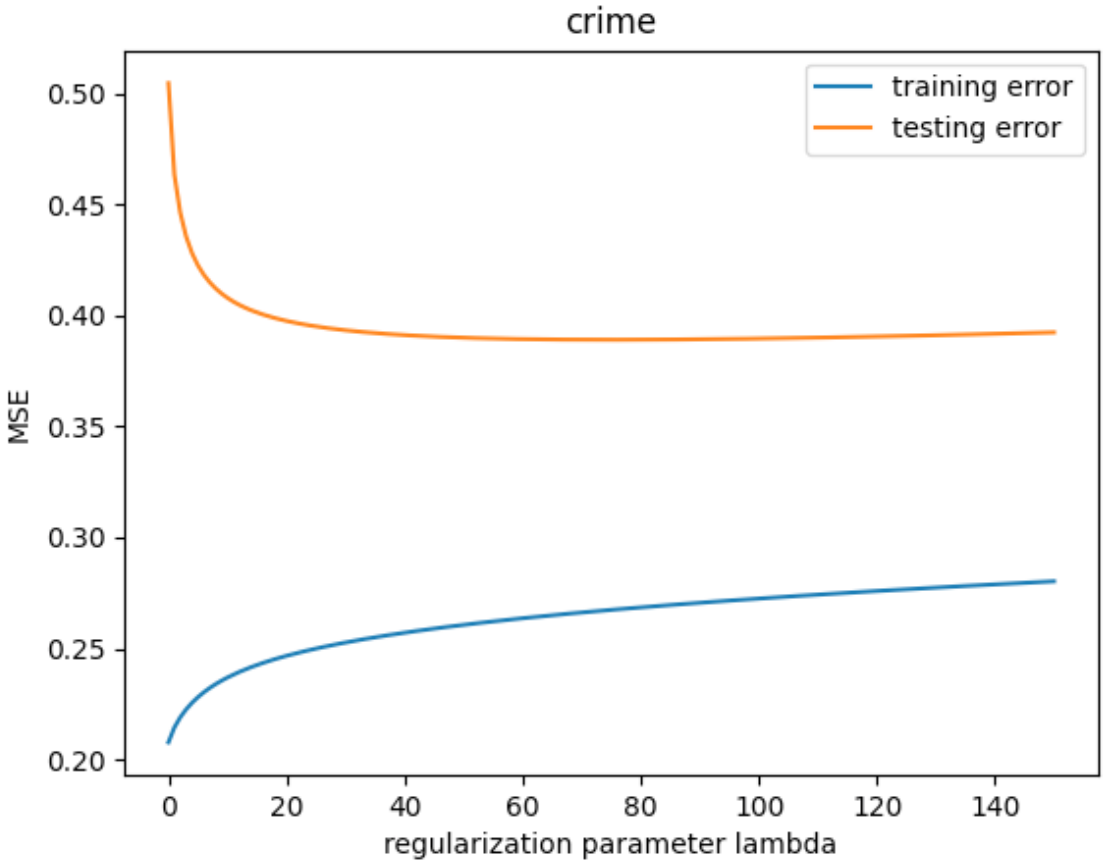
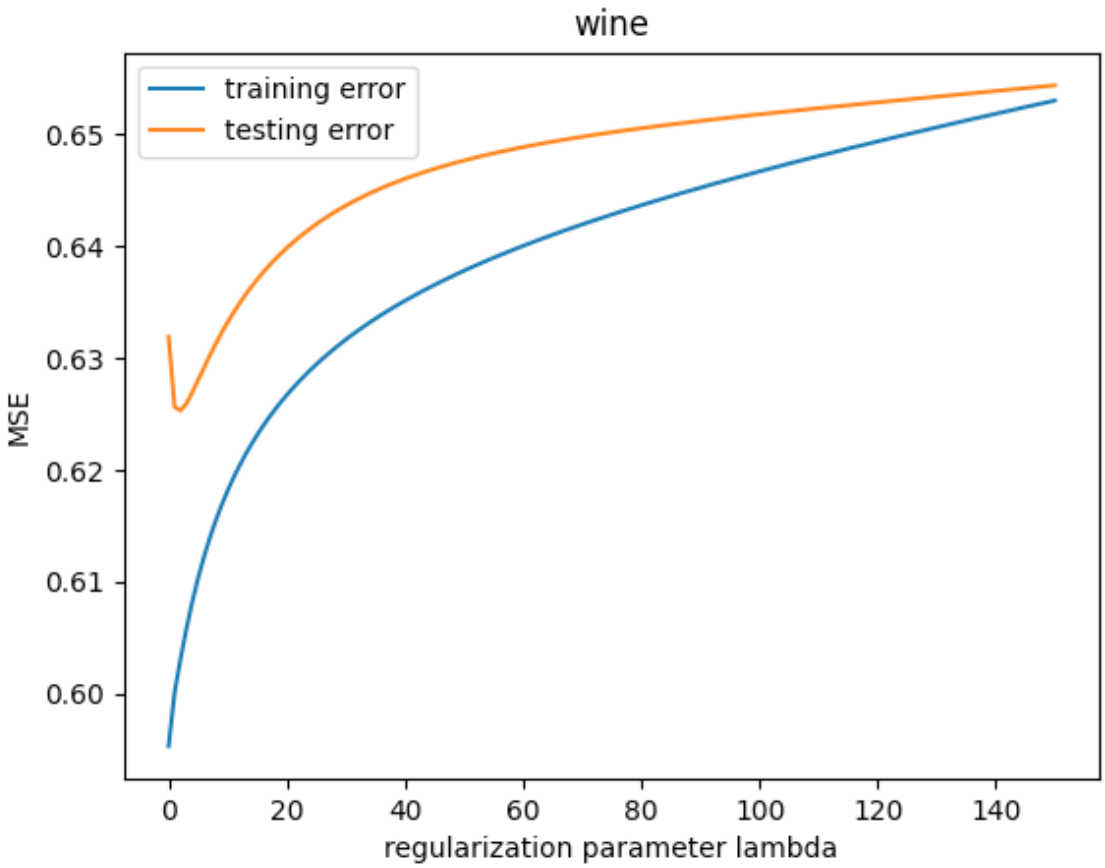
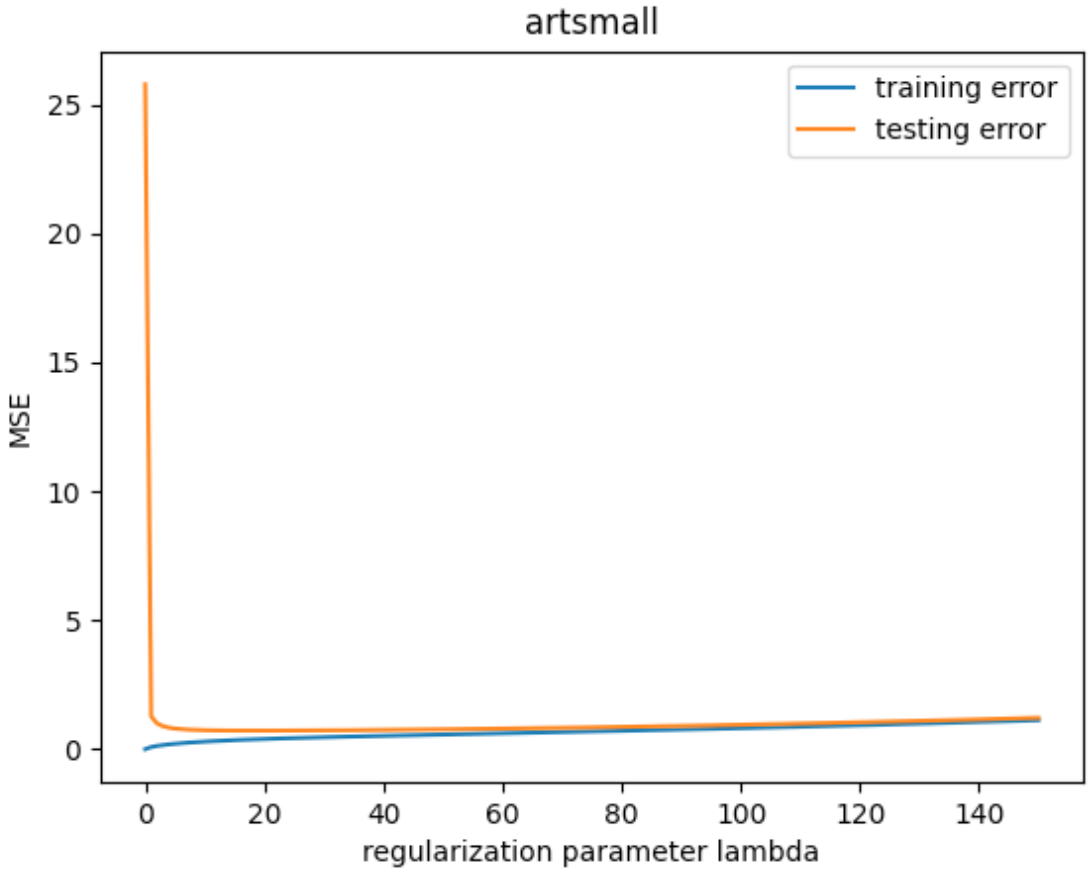
PP2 Report
Brendan McShane

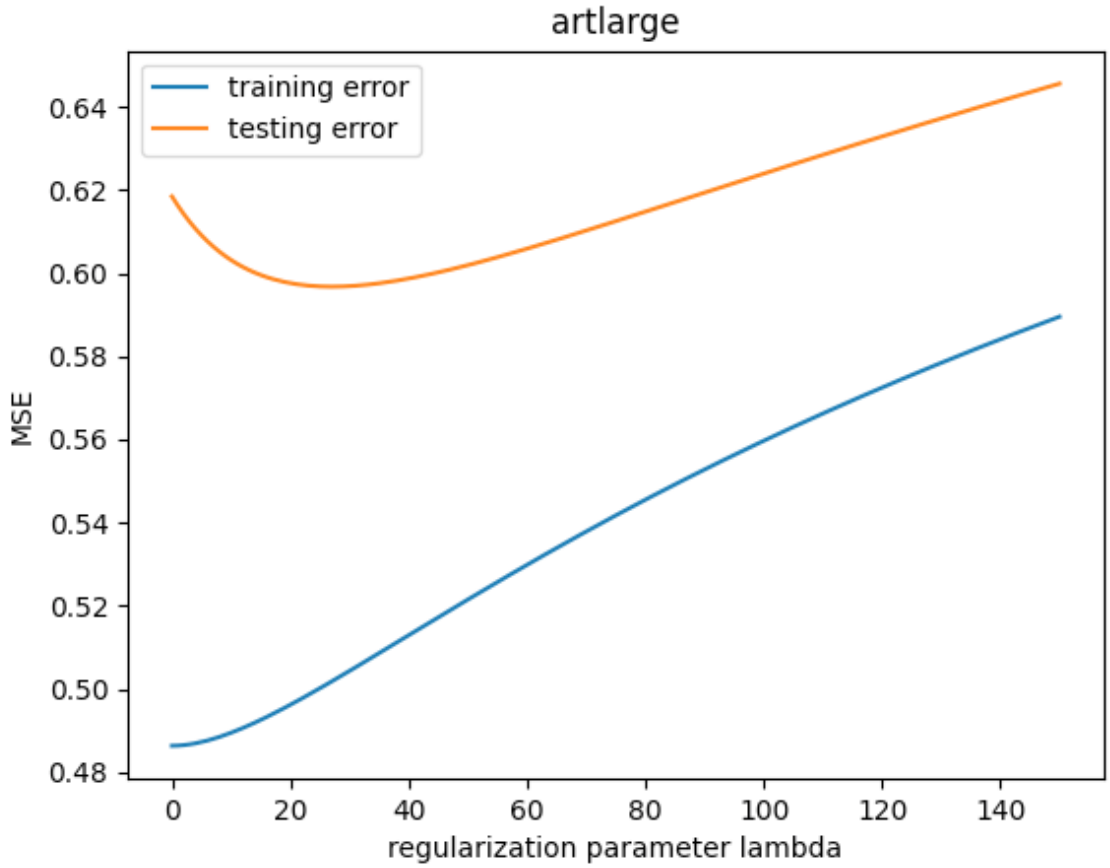
Part 1
The results for part one are copy and pasted below for clarity, they're graphs depicting the size of the regularization parameter lambda vs MSE on the testing and training sets. For the crime dataset, training MSE starts extremely low with less regularization and then quickly rizes and plateaus with a higher lambda value. This makes sense because a larger lambda fights against model complexity more and forces the solution to generalize better to data outside our training set, instead of just training to perfectly predict data points we've already seen. This concept is also reflected in the testing error, as it starts very high with a low lambda value and then shoots down and plateaus and the model generalizes more and more to fit data outside the training set. A very similar phenomenon occurs with the artsmall dataset, except the error for the training set is much flatter and the error for the testing set shoots down much more sharply and flattens out quicker. I think this is due to the nature of the artificial dataset, because the testing and training MSE both seem to zero in on the MSE of the hidden true function that is generating the data (MSE=.533). The same thing isn't happening in our results for the artlarge dataset, however. Maybe that's because the data for the training and testing datasets are very similar and thus the more we regularize on the training set, the worse our model fits the training data and therefore fits the testing data. Both MSEs see a consistent increase as lambda increases. I think that's the same thing that's going on in the wine dataset, the training and testing data are so similar that regularization actually works against us.

To answer the questions at the end of Part 1, we can't use the training set MSE to select lambda because the entire point of lambda is to reduce model complexity to increase performance on data outside of our training set, so that's the metric we care about. If we used the MSE from the training set, we wouldn't regularize at all and our model would just fit the training data perfectly and perform terribly on any outside data. Generally, a higher lambda will mean a lower testing MSE because the model will generalize better to data outside the training set. This isn't the case for two out of the four datasets, however, and I've explained why I think that is above.


crime
training error
testing error
MSE
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0
20
40
60
80
100
120
140
regularization parameter lambda


wine
training error
testing error
0.65
0.64
0.63
0.62
0.61
0.60
MSE
0
20
40
60
80
100
120
140
regularization parameter lambda


artsmall
training error
testing error
MSE
25
20
15
10
5
0
0
20
40
60
80
100
120
140
regularization parameter lambda



Part 2

```
Part 2
crime
{'best lambda': 150, 'associated mse': 0.35001906956796647, 'associated runtime': 1.11216402053833}
wine
{'best lambda': 2, 'associated mse': 0.6459688229371905, 'associated runtime': 0.27094030380249023}
artsmall
{'best lambda': 18, 'associated mse': 0.7215067060739677, 'associated runtime': 0.8775889873504639}
artlarge
{'best lambda': 22, 'associated mse': 0.5882896773201547, 'associated runtime': 2.599018096923828}
```

Above are my results for part 2. I've been told my lambda for the crime dataset is off the mark and I couldn't figure out why in time, but the other lambdas should be approximately correct. I believe the actual correct lambda for the crime dataset should be around ~70, as that is where the MSE begins to plateau for the testing set. For the wine dataset an optimal lambda of 2 makes sense, because that's where we see the initial dip in the testing set MSE for this data and it should be approximately the global minimum. The MSE is slightly higher here than it is in part 1. That might be a result of it being averaged over 10 folds and potentially being more accurate than the estimate in part 1. Maybe the fact that for each fold we're only working with 90% of the data points results in our MSE being a little higher. For artsmall an optimal lambda of 18 seems to make sense, although any value in the general area would seem to get the job done. The graph is relatively flat at that point, and the testing MSE isn't going any lower than

that. Although it's a little hard to tell, the MSE also seems to lign up with the graph for part 1. For artlarge, the ideal lambda is 22 which is around where the minimum is for the graph MSE vs lambda. The part 2 MSE is a little lower than the part 1 MSE, which again I'll hypothesize is a result of it being averaged over 10 folds and being more accurate.

Part 3

```
Part 3
crime
{'alpha': 425.6453346429579, 'beta': 3.2504320906460213, 'lambda': 130.95038529427057, 'MSE': 0.3911023051920137, 'time':
0.0893089771270752}
wine
{'alpha': 6.163876730705895, 'beta': 1.609809320070937, 'lambda': 3.828948344288553, 'MSE': 0.6267461126898939, 'time': 0.
004992961883544922}
artsmall
{'alpha': 5.154668973370687, 'beta': 3.1542650922996014, 'lambda': 1.6341901592084327, 'MSE': 1.0634956452428102, 'time':
0.05190706253051758}
artlarge
{'alpha': 10.285792799507663, 'beta': 1.8603093607409482, 'lambda': 5.529076516290227, 'MSE': 0.6083085989532367, 'time':
0.03433394432067871}
```

For crime, the ideal lambda aligns with our results from part 1 because it's along the plateau where MSE flattens out and the MSE is spot on. For wine, the lambda and MSE are also in line with what we've seen from part 1. For artsmall the lambda is right about where the curve sharply plateaus for testing MSE, and the MSE reflects where that happens on the y-axis. For artlarge, the lambda seems to undershoot what we saw in part 1, it's 5.529 instead of somewhere in the range of 20 to 30. The MSE seems accurate for the chosen lambda.

Part 4

With respect to MSE, cross validation appears to perform slightly better on the crime and artlarge datasets and significantly better on the artsmall dataset. The lambdas chosen seem to be larger in general than those chosen by the Bayesian approach, except for the wine dataset where the Bayesian approach chooses a slightly larger lambda and performs better. The Bayesian approach also appears to be notably faster than cross validation, which makes sense because with CV we have to iterate over 10 folds for each dataset. From these results it would appear that cross validation performs slightly better and chooses larger lambdas, while the Bayesian approach is faster and more efficient. Although, my results for the crime dataset in the CV approach are off, so maybe the actual MSE is higher than what is seen in the part 2 results. This would imply that the Bayesian approach performs better on real world datasets with larger lambdas and more regularization, while the cross validation approach performs better on generated data where the testing data is very similar to the training data, and a more complex model with less regularization (and a smaller lambda) performs better. This would also explain why the CV approach performs slightly better on the artlarge dataset and significantly better on the artsmall dataset. With less regularization, the CV approach can overfit on the training data for artsmall more efficiently, which should be fine because of how similar the testing data is.