Seqian Wang, 260377179
In collaboration with Sulantha Mathotaarachchi and Maxime Parent

Single-layer classification

The learning rate affects how receptive a person or system is to feedback. There is a trade-off between a small rate and a high rate, and neither extreme is optimal. Indeed, a small learning rate will take a long time to train (need a higher number of iterations), while a big learning rate will constantly overshoot and undershoot the weights, failing to find the right level. Note: For the sake of time, I only ran the experiment 5 times for a given learning rate. Having more trials per learning rate would yield more stable results for the error. In other words, I would expect to find the peaks with values around 10 to be lessened in the following figure and the standard deviation reduced.

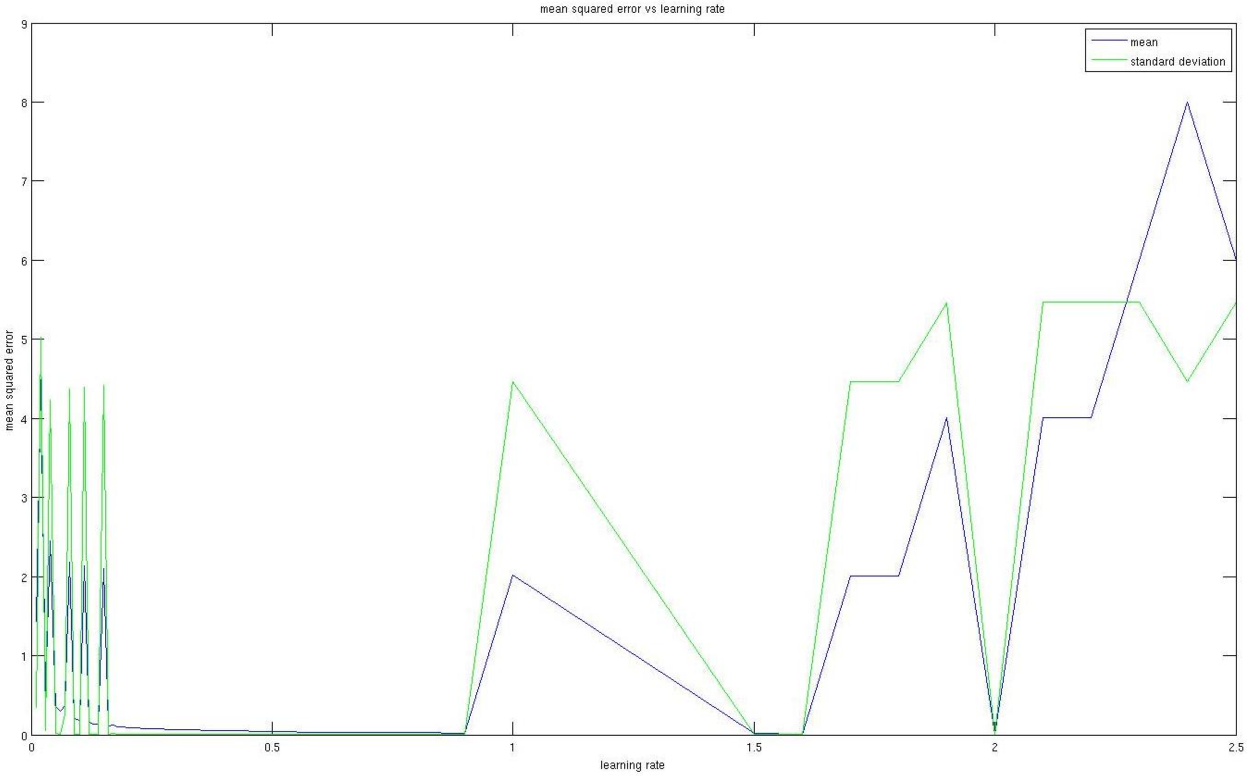


Figure 1: MSE vs Learning Rate. The following learning rates have been tested and used as data points: [0.01:0.01:0.2 0.3:0.1:1 1.5:0.1:2.5]. Since we are using 500 trials, the error at a learning rate of 0.01 is barely noticeable, and I would expect a larger error if less trials were used. As for larger learning rate values, there are more instances where the classifier fails (high error). 0.8 seems most optimal from this figure.

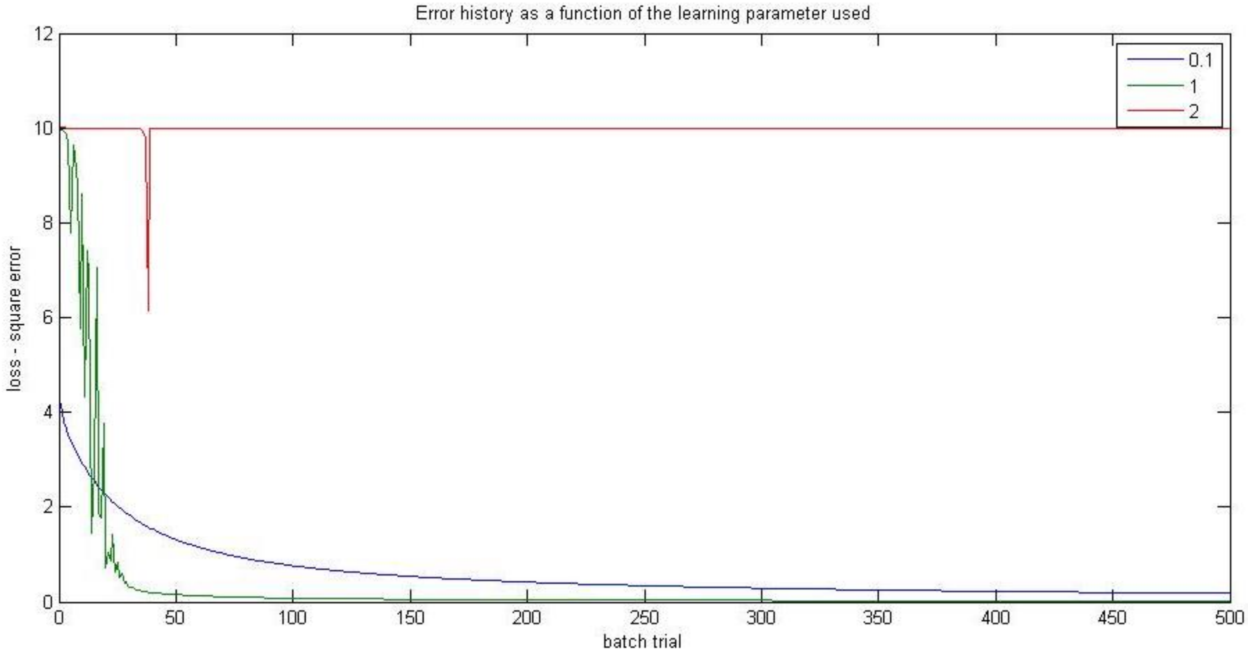
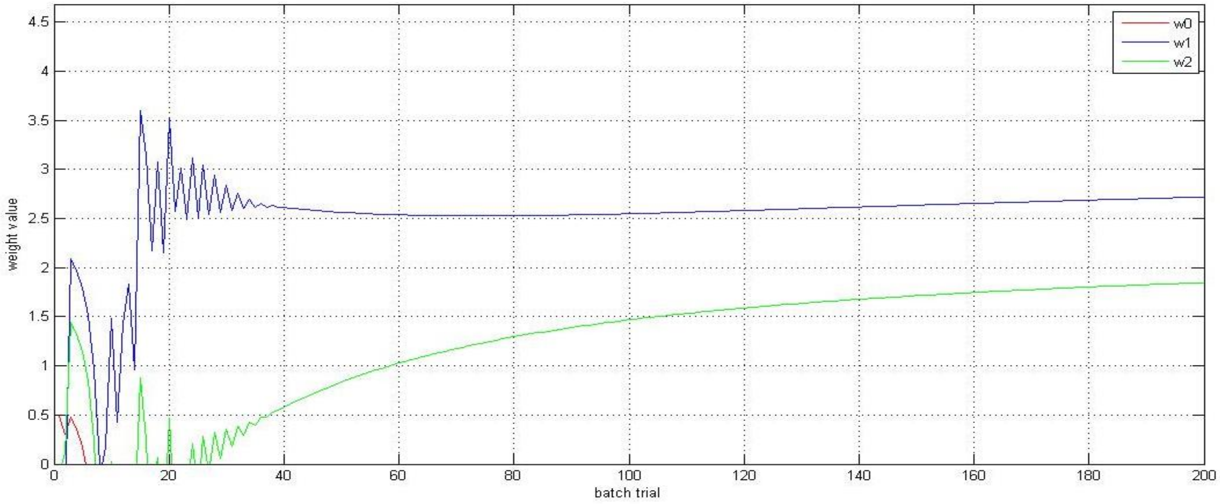


Figure 2: Error history as a function of the learning parameter. These present stereotypical instances of the instances I observed. With a low learning rate, it takes a lot of trials to reduce the error. With eta=1, there are a lot of fluctuations for the error, but it reduces the error very fast. Finally, with a high eta (2), the classifier simply fails to learn.

Single-layer classification with Regularization

For the following, we used a learning rate of 0.8 and 200 trials per run, as it was found to be most optimal in the previous exercise (least MSE).The regularization parameter defines the amount of decay. It is a way to prevent over-fitting and to limit large weight values. Here below is an error assessment for different values of regularization parameters. We tested the MSE as a function of the regularization parameter. We generated our data with 200 batch trials per run, 10 runs per regularization parameter, a learning rate of 0.8, and the following regularization parameters: [0.01:0.02:0.1 0.2:0.1:1].



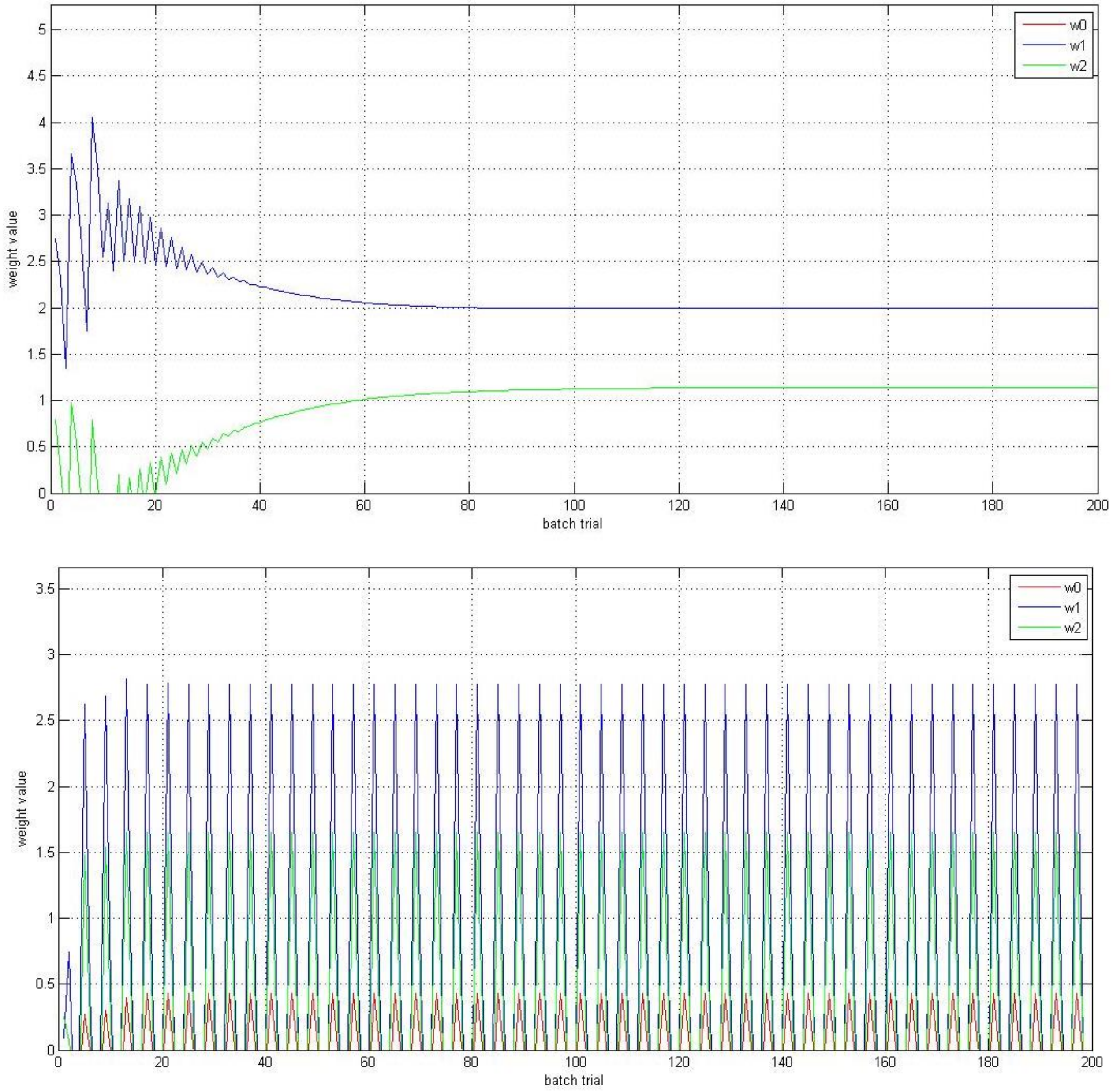


Figure 3: From top to bottom, weight values vs iteration number using the hyperparameter 0.001, 0.01 and 1 respectively. w0 is red, w1 is blue, and w2 is green. The 0.001 instance takes more time to learn and to find the optimal weight values. In contrast, the 0.01 instance stabilizes its weight values at around trial 100. Finally, the 1 instance keeps overshooting and undershooting the weight values, going into a sine wave and fails to converge towards a weight value.

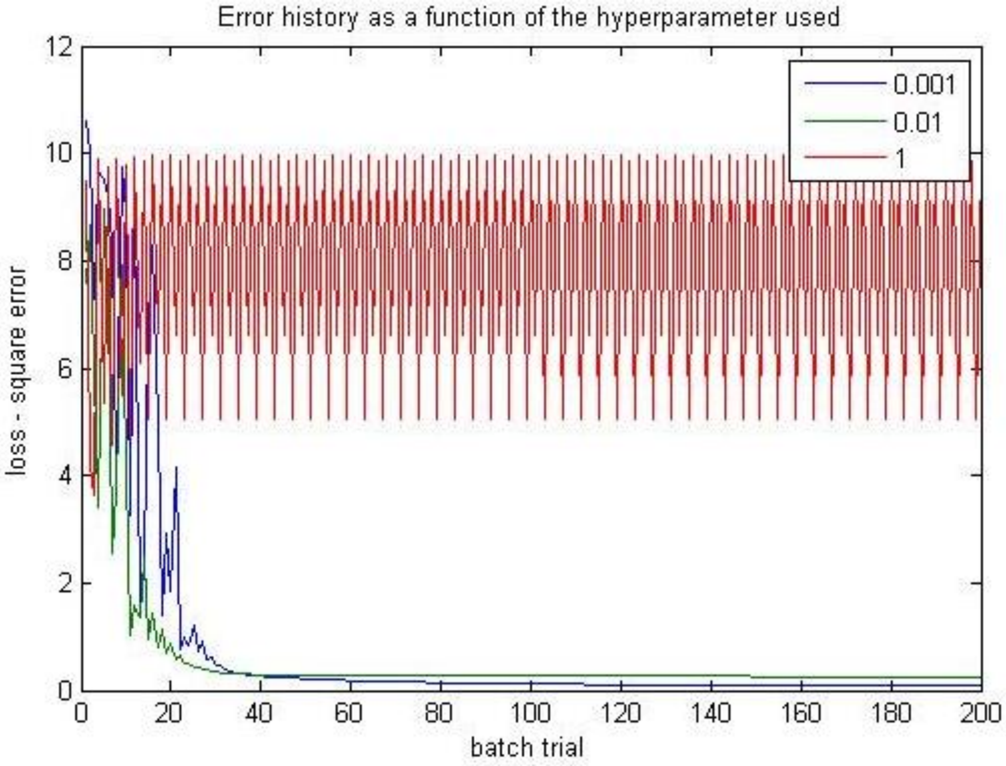


Figure 4: The error history for different regularization parameter values. The 0.01 reaches the lowest MSE the fastest, although the 0.001 instance manages to reduce the error even more in the long run. The 1 hyperparameter fails to reduce the error and keeps going back and forth (just like for the weight values).

Figure 4 and 5 show stereotypical examples of the classifier behaviour for different regularization parameters. From the presented data, we can affirm that the 0.001 hyperparameter value for a learning rate of 0.8 and 200 batch trials is the best (least MSE), but that a 0.01 hyperparameter could be more useful in a case where batch trials are limited (ex.: 20 trials).

Receptive Field Estimation using Regression

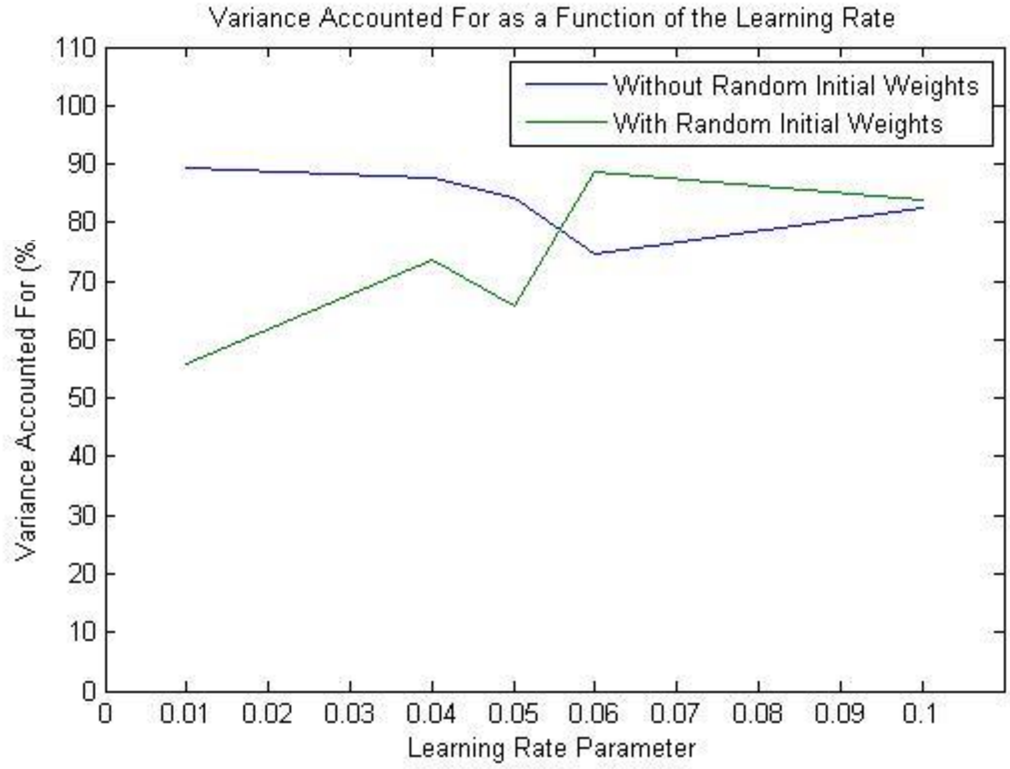


Figure 5: The variance accounted for as a function of the learning rate [0.01 0.04 0.05 0.06 0.1], using 50 batch trials.

Although not very informative here due to the lack of data points and processing power (I am running this on a tablet/netbook), we can use a similar figure as the one above to determine the optimal learning rate parameter (with the highest VAF). I would expect high values of eta (>0.1) to have low VAF because of overfitting problems. Likewise, with a low number of batch trials, I would expect instances with low learning rate parameters to have low VAF, as they would be slow to learn the model. From observations, I would expect a higher VAF for runs with initial zero weights vs random weights, especially for runs with a low number of trials. This is because runs without random weights has less variability and has a better initial fit (many data points for the receptive field is at level 0).

Receptive Field Estimation using Regression and Regularization

For the following exercise, 50 trials per run and a learning rate of 0.05 have been used.

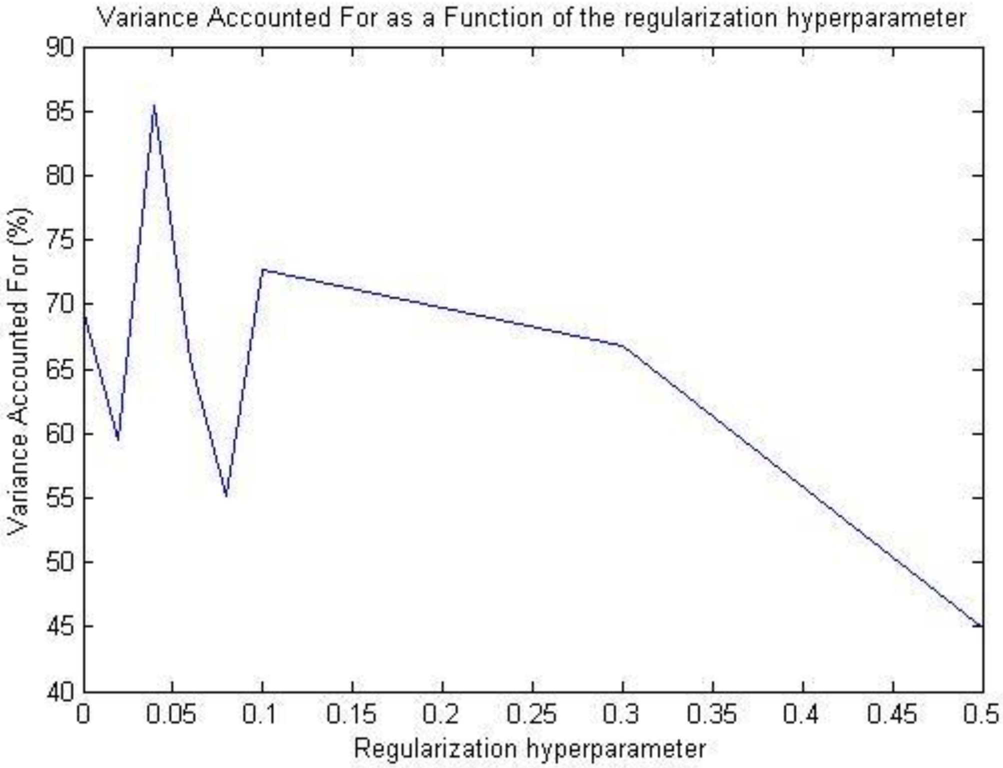


Figure 6: The variance accounted for as a function of the regularization parameter [0:0.02:0.1 0.3 0.5], using 50 batch trials, but only 1 run per regularization parameter.

Again, I am short of data points, time and processing power, and so the figure is not too informative. More runs per alpha value would also yield a more consistent plot (as in we can see more of a trend or relation). Nevertheless, we can determine the optimum regularization hyperparameter with a figure like the one above (here, 0.05). From observing the estimated kernel, I believe that a high alpha value will constantly try to keep the kernel in check, and the simulation will constantly overshoot and undershoot. In contrast, a reasonable regularization hyperparameter will help the estimated kernel reach a good VAF level with less batch trials than if without regularization.

The predicted vs measured output gives a sense of how correlated and close the observed and estimated values are. In that sense, a higher VAF means a higher correlation between observed and expected (as we want), and vice-versa.