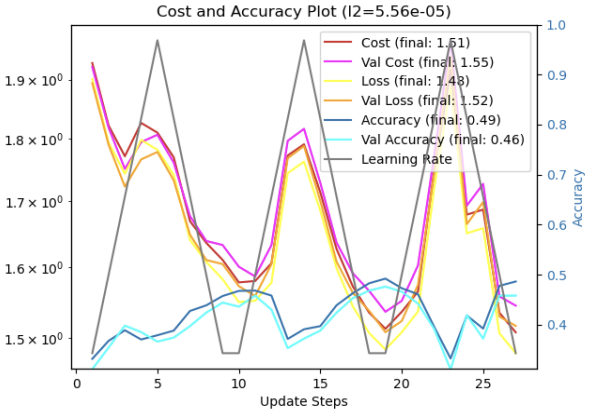
A2 Report Bonus - Kevin Armbruster

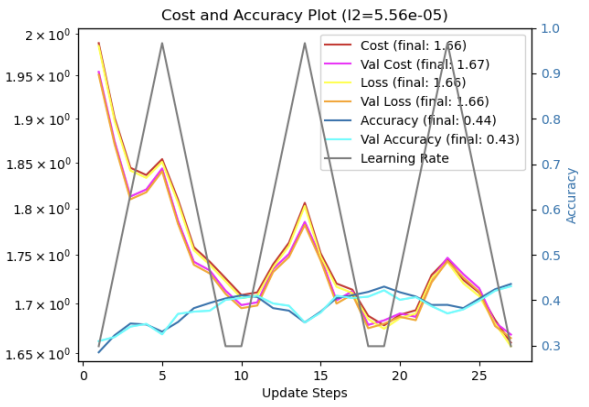
Improvements

1. Data Augmentation
    a. The data augmentation is randomly flipping or rotating the image in either direction.
    b. I tested my improvements with my previously best network and the results for 3 cycles diminished slightly to 48% of previously > 50%.
    c. This is probably due to the higher complexity of the training problem, that might not be necessary during test time. A longer training duration might offset this observation.
2. Inverted Dropout
    a. Inverted Dropout also slightly diminished the accuracy results of my best previous network to 46-47% of previously 48%.
    b. As I observed already in the standard report, all the good regularization strengths are fairly weak, hence my take is that the additional dropout regularization is just too much.
    c. A longer training duration could improve the outcome, because I noticed that the general trend of the accuracy is upwards after 3 cycles.
3. Investigating a much larger network
    a. I tested a hidden layer size of 500 trying out different regularization strengths, while also using the 2 previously mentioned improvements.
    b. I noticed that the training time significantly increases and the additional neurons either slightly increase the accuracy to around 49-50% or heavily diminish the accuracy of the network with an unfitting regularization.

Here I have the standard cost, loss, accuracy curves for a network with these improvements and 500 (left) vs 50 (right) neurons in the hidden layer.

Final test accuracy for 50 nodes: 0.4274
Final test accuracy for 500 nodes: 0.4956





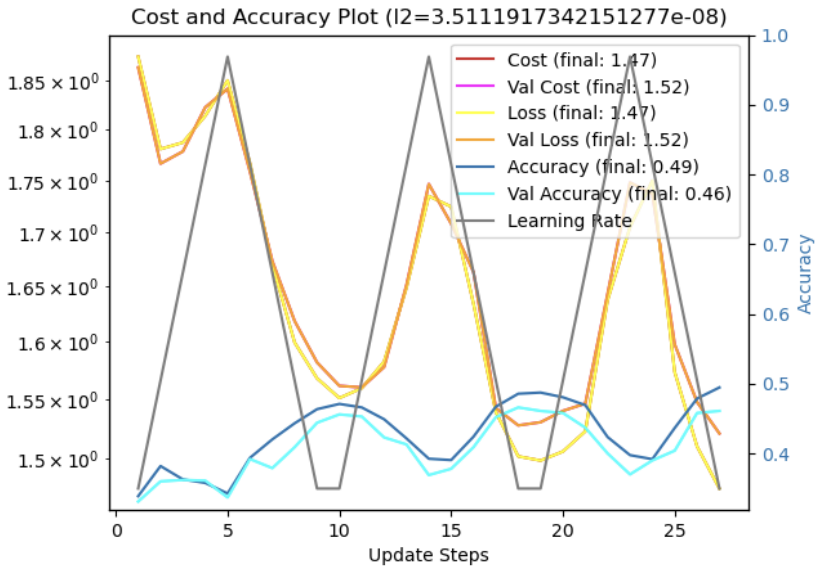
Optimization

I did a gridsearch for the best L2 strength using 3 cycles, batch size of 100, eta_min= 1e-5, eta_max= 1e-1, eta_step_size= 500, 50% dropout chance and data augmentation. (Similar to the default report, but with the improvements.) Additionally, I investigated 3 vs 6 cycles in both networks.
We observe that the networks in my setup performe better with on the final testset using either a wider architecture or more cycles. The spikes in either direction of the loss/cost curves during the learning rate cycle are bigger in the wider 5000 node network which indicates a more insatiable training process. Given this investigation took close to 2 days of computational time, I would argue that the additional accuracy has to be weighed against the computational effort.

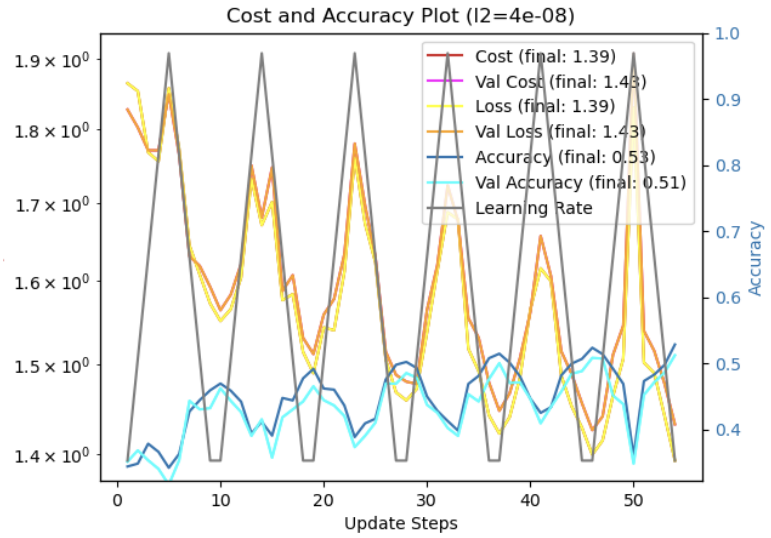
500 nodes - L2:

1. 0.00000004: 0.485
2. 0.00284804: 0.48
3. 0.00000152: 0.479
4. 0.00006579: 0.479
5. 0.00001874: 0.478
6. 0.00023101: 0.478
7. 0.00000001: 0.471
8. 0.00081113: 0.467
9. 0.00000534: 0.465
10. 0.00000043: 0.464
11. 0.00000012: 0.456
12. 0.01000000: 0.451

Final test accuracy after 3 cycles = 0.4931



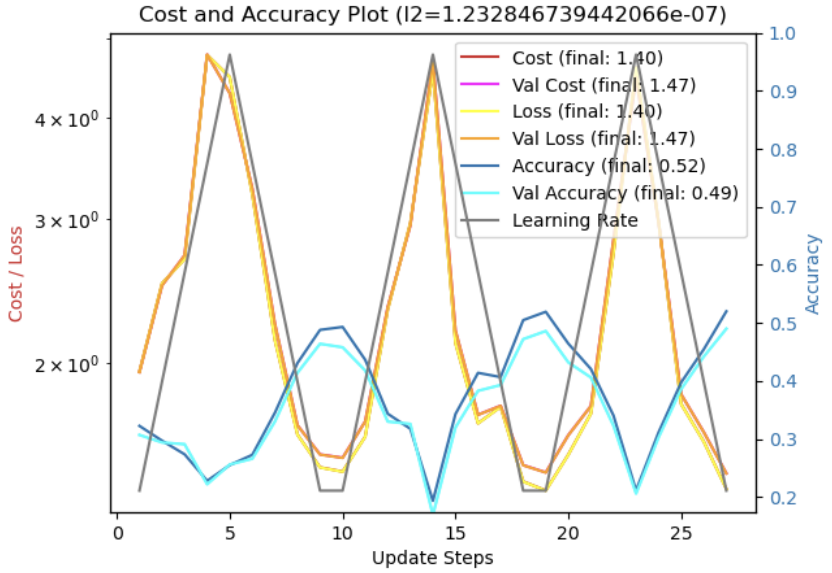
Final test accuracy after 6 cycles = 0.5084



5000 nodes - L2:

1. 0.00000012: 0.498
2. 0.00000004: 0.494
3. 0.00006579: 0.493
4. 0.00000001: 0.489
5. 0.00000043: 0.486
6. 0.00000152: 0.479
7. 0.00284804: 0.475
8. 0.00081113: 0.472
9. 0.00000534: 0.471
10. 0.00001874: 0.468
11. 0.01000000: 0.463
12. 0.00023101: 0.455

Final test accuracy after 3 cycles = 0.4989



Final test accuracy after 6 cycles = 0.5186