

A. Supplemental Materials

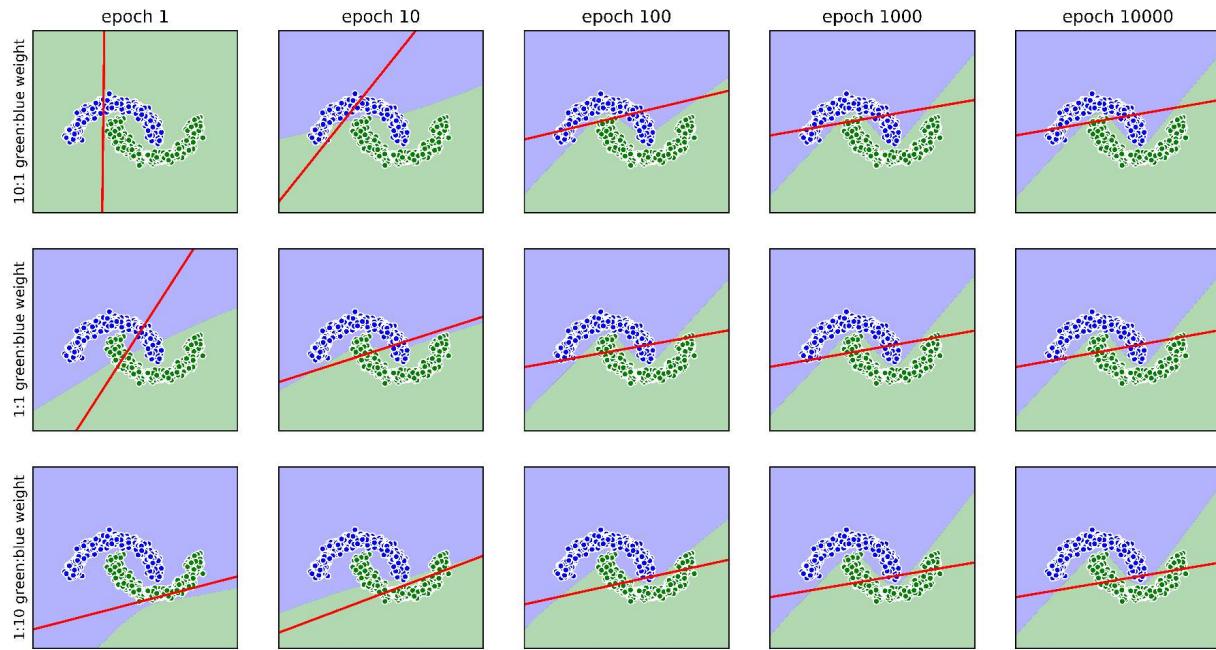


Figure A.1. Results on non linearly-separable moons dataset.

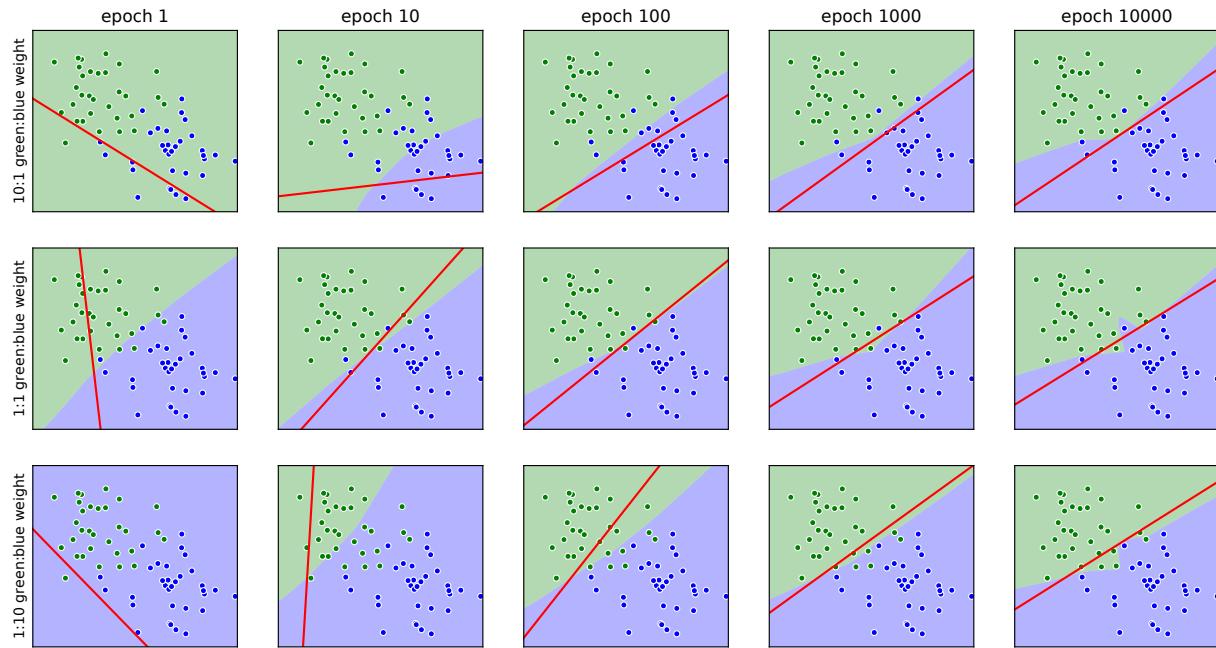


Figure A.2. Results on overlapping Gaussians dataset.

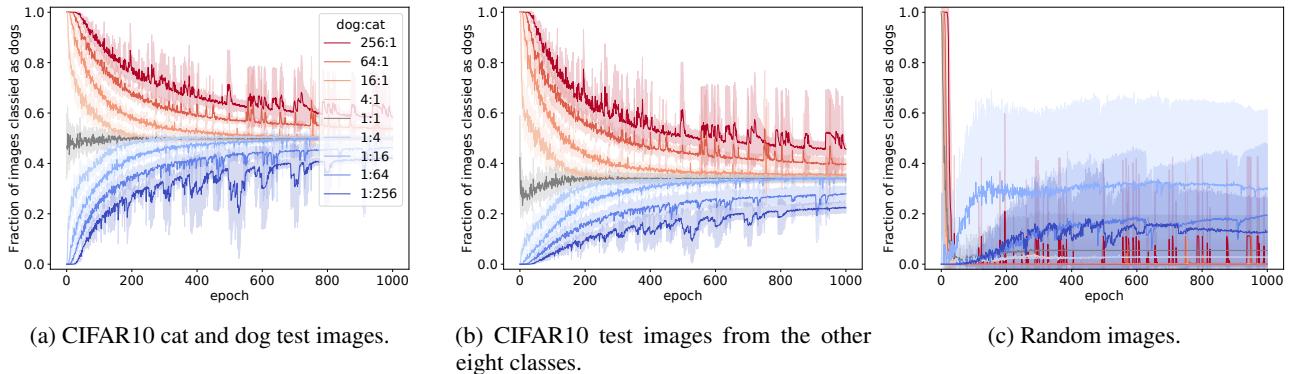


Figure A.3. Effect of early stopping on L2 regularized models. Plots are structured as in Figure 4.

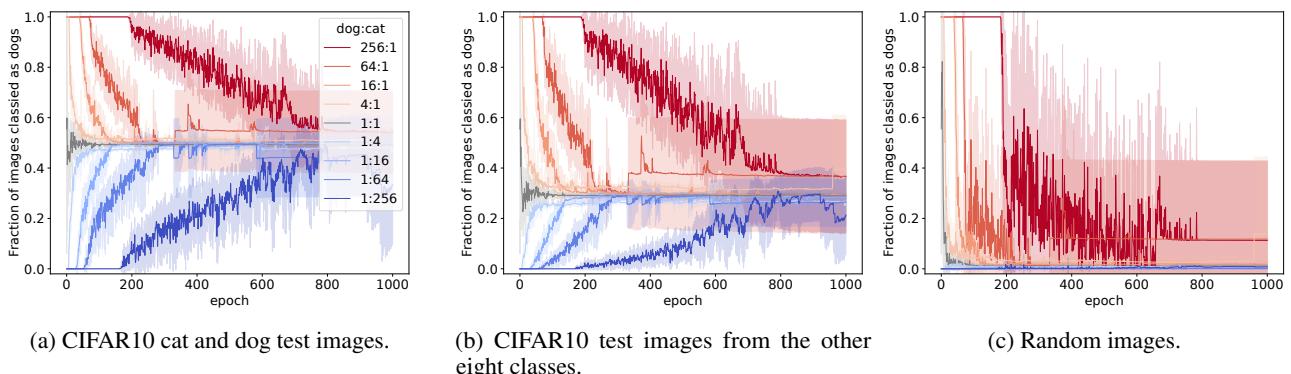
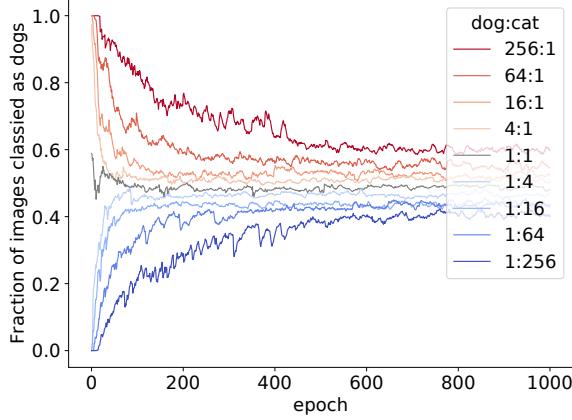
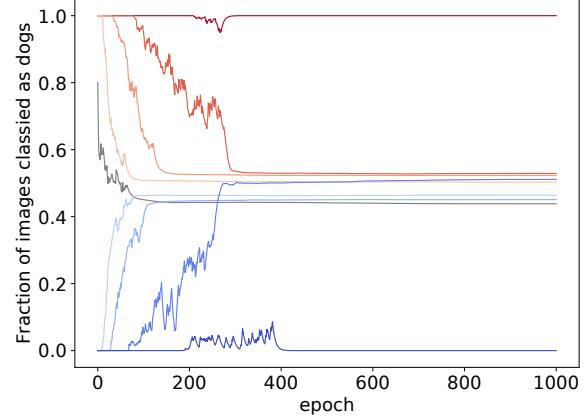


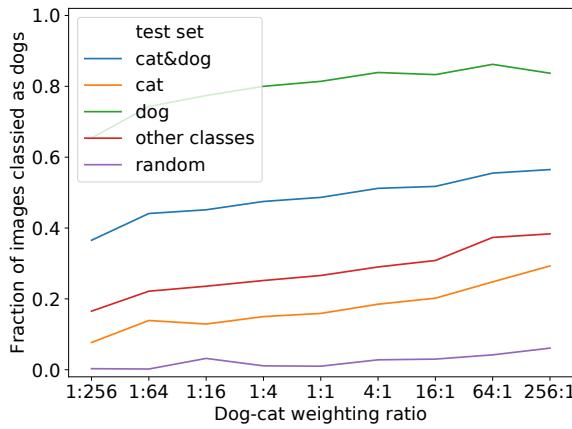
Figure A.4. Effect of early stopping on models with dropout.



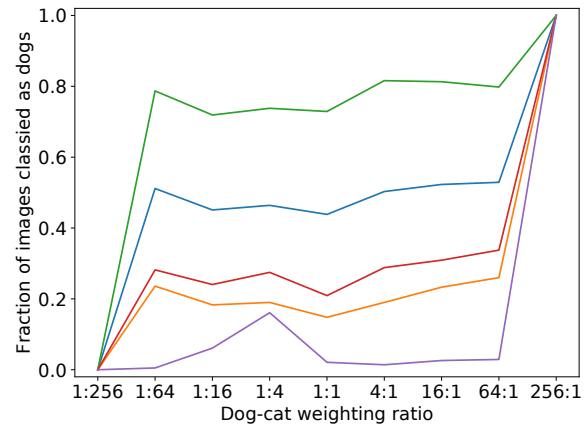
(a) ResNet with batchnorm.



(b) ResNet without batchnorm.

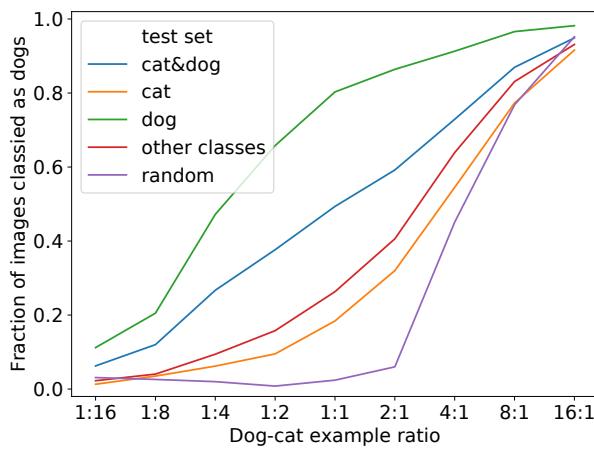


(c) ResNet with batchnorm.

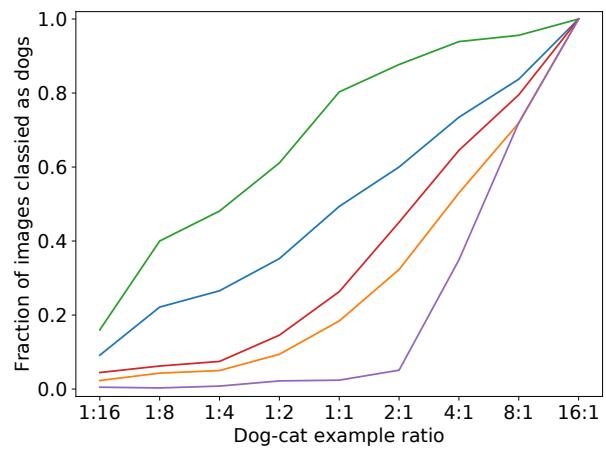


(d) ResNet without batchnorm.

Figure A.5. Relationship between early stopping and importance weighting (a,b), and final classification ratios vs. importance weighting (c,d) for ResNet models on CIFAR with and without batch normalization. Plots are structured as in Figure 4.



(a) Without rebalancing via importance weighting.



(b) With rebalancing via importance weighting.

Figure A.6. Attempting to correct for class imbalance with importance weights. Results are shown for different sub-sampling ratios (X-axis). Our results suggest that while class balance at training time impacts the classification ratios, IW-ERM does not.

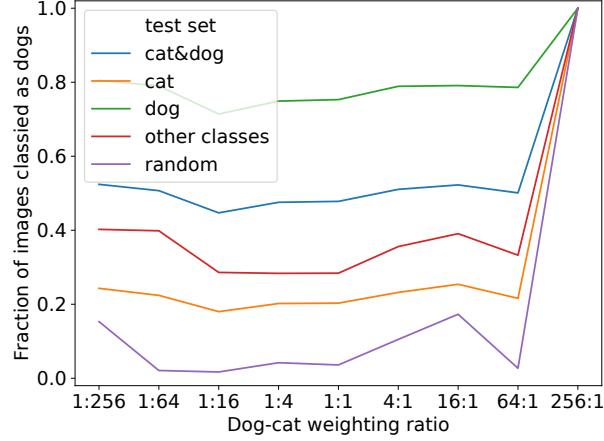
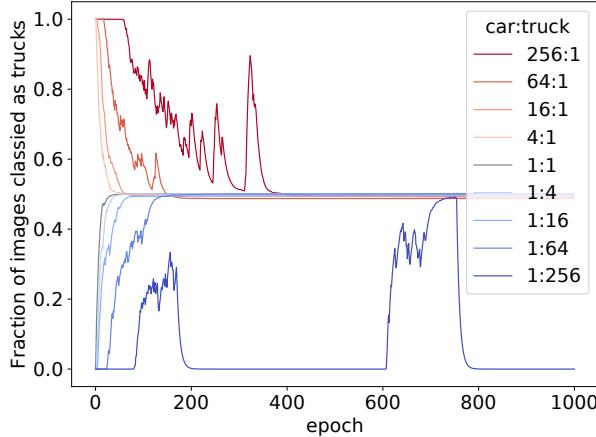
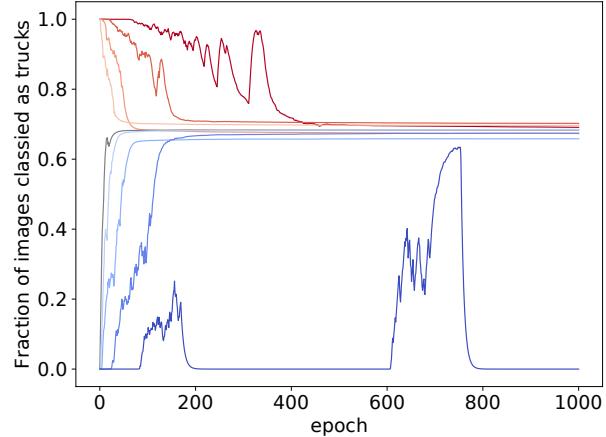


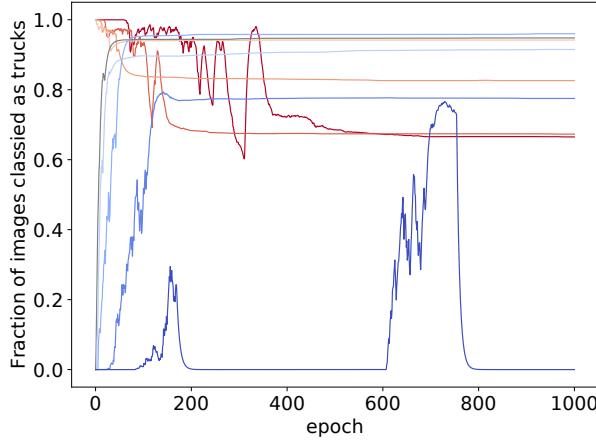
Figure A.7. Classification ratios vs importance weights on CIFAR10 where the labels of 5% of each class in the training set are flipped. This experiments shows that even when the true classes are inseparable, importance weighting has surprisingly little effect so long as the neural network can separate the training data.



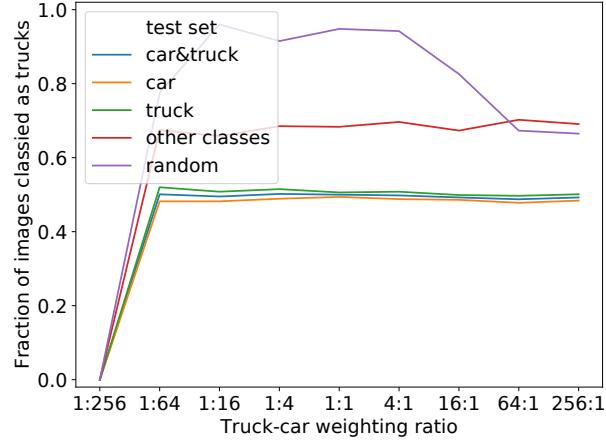
(a) CIFAR10 car and truck test images.



(b) CIFAR10 test images from non-car/truck classes.

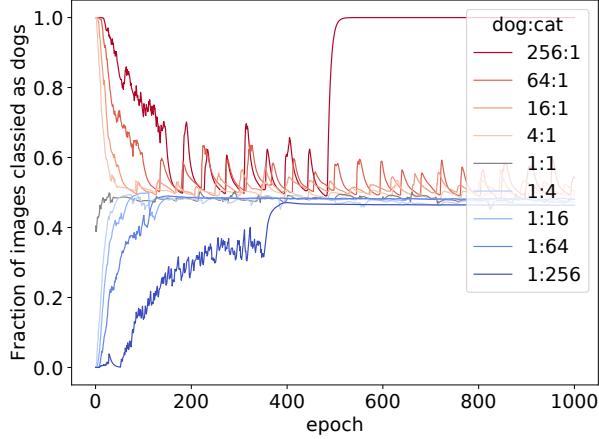


(c) Random images.

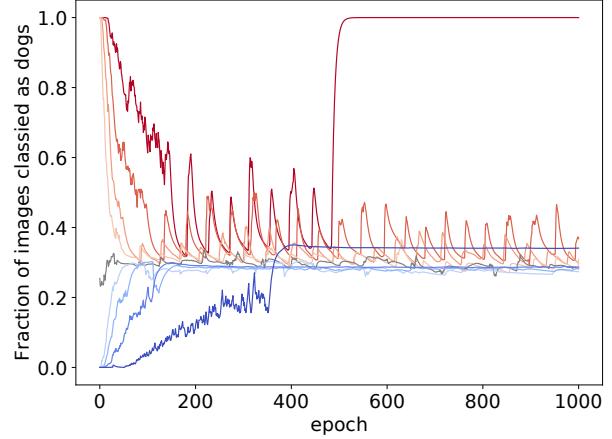


(d) Classification ratios after training on car/truck images.

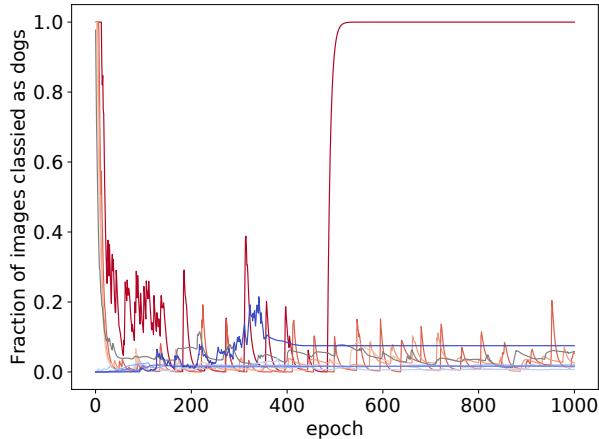
Figure A.8. Results for training on automobile and truck classes from CIFAR10. Same setup as 4, but without standard deviations over multiple runs.



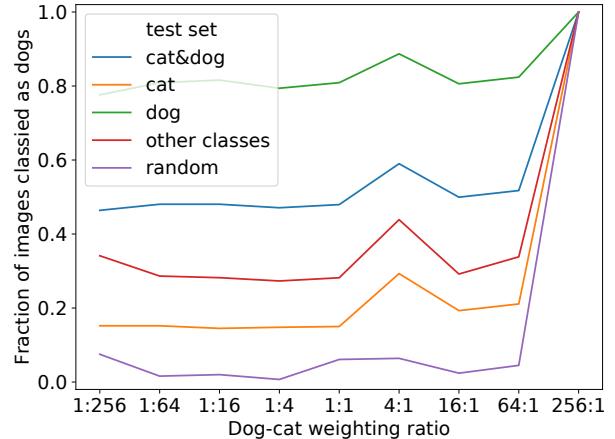
(a) CIFAR10 cat and dog test images.



(b) CIFAR10 test images from non-cat/dog classes.



(c) Random images.



(d) Classification ratios after training on cat/dog images.

Figure A.9. Results from training a convolutional network with the Adam optimizer with learning rate $1e - 4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$. The setup and all other model hyperparameters are the same as in 4, but without standard deviations over multiple runs.