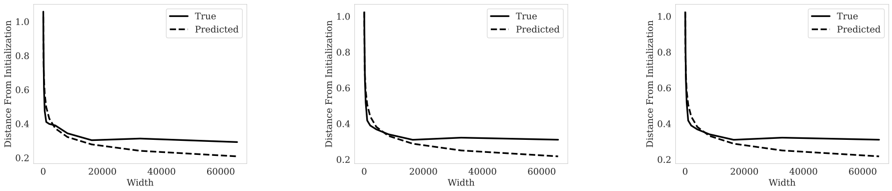
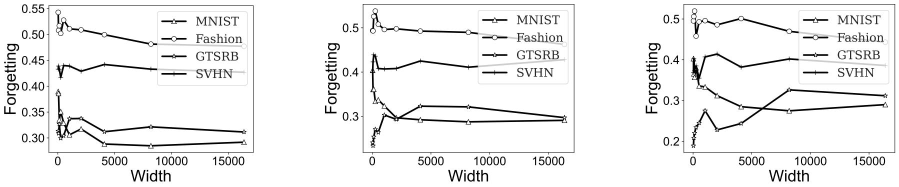


(a) .1 Dropout Probability

(b) .2 Dropout Probability

(c) .3 Dropout Probability

Figure 1: Distance from Initialization for with Dropout for different Dropout probabilities.



(a) 4 Hidden Layers

(b) 5 Hidden Layers

(c) 6 Hidden Layers

Figure 2: We plot forgetting as width of the network is increased for different number of hidden layers.

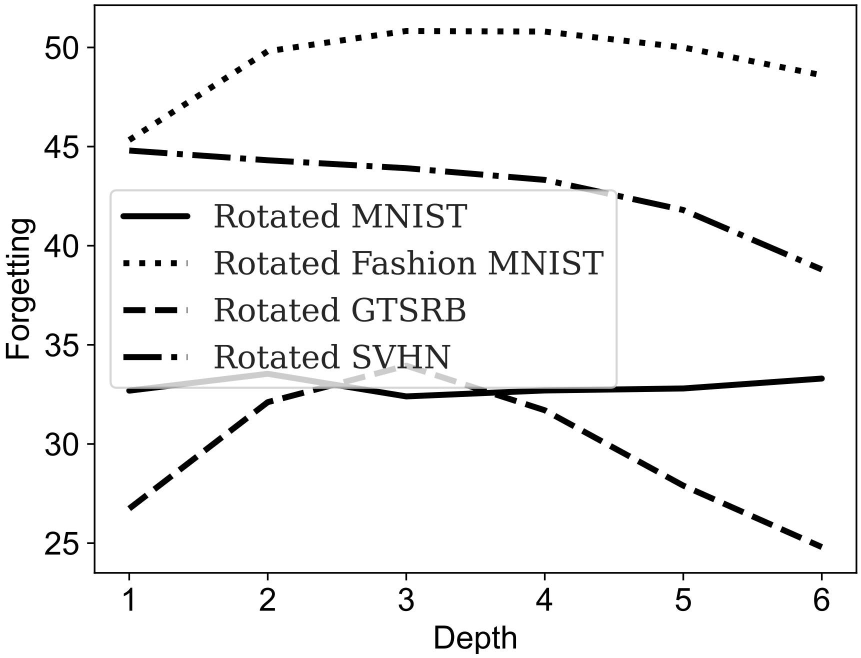
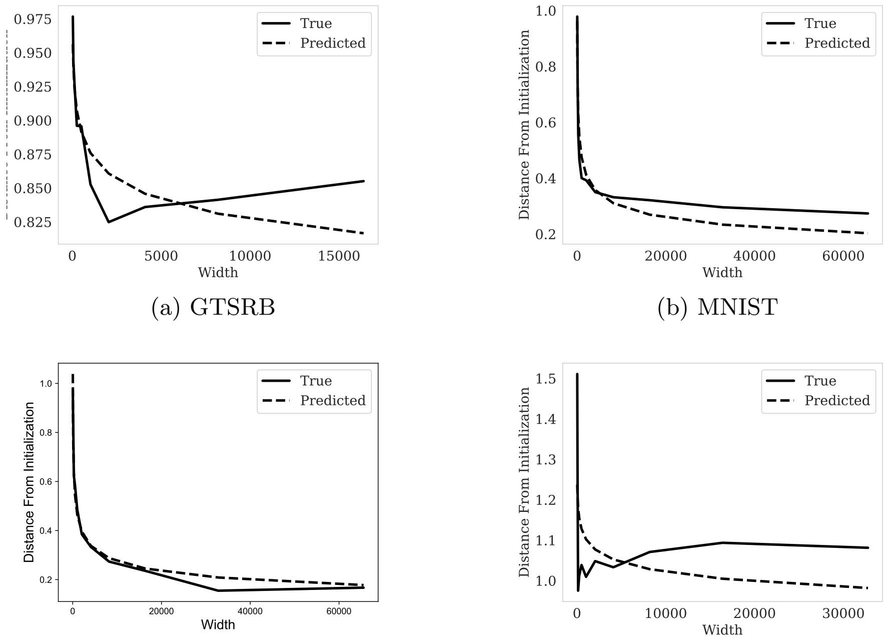


Figure 3: Average Forgetting as Depth is increased. We see that as depth is increased, forgetting increases. However, as depth is increased further, the accuracy goes down due to vanishing gradients. This artificially causes the forgetting to decrease.



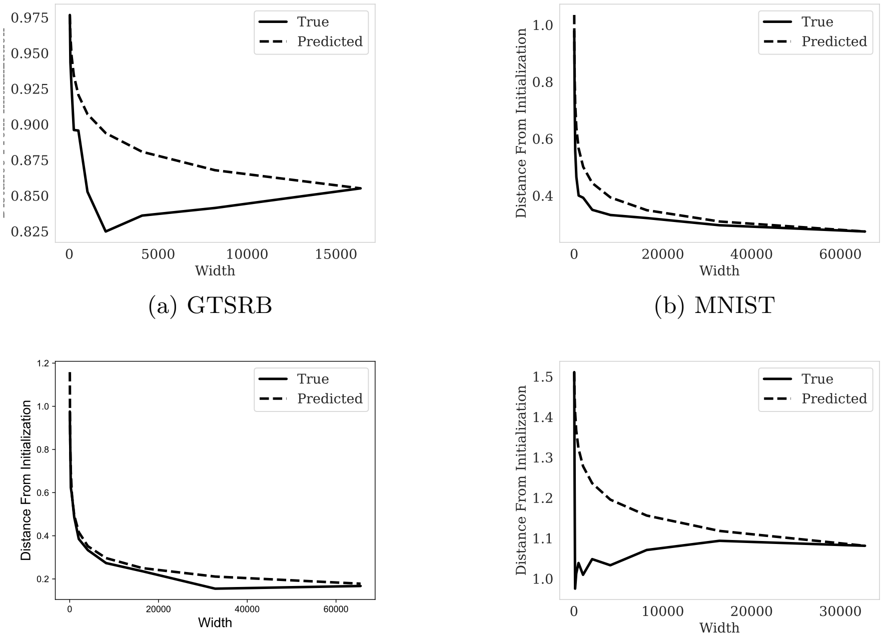
(a) GTSRB

(b) MNIST

(c) Fashion MNIST

(d) SVHN

Figure 4: We plot the distance from initialization for training at different widths on different datasets. In the "Predicted" line, we plot the β and γ values (from Assumption 4.3) that best fit the curve. Here, we do not require that the predicted curve upper bounds the true curve as in Assumption 4.3 for visualization purposes.



(a) GTSRB

(b) MNIST

(c) Fashion MNIST

(d) SVHN

Figure 5: We plot the distance from initialization for training at different widths on different datasets. In the "Predicted" line, we plot the approximation from Assumption 4.3. where the predicted curve does indeed upper bound the true curve as in Assumption 4.3.