We notice that the attack which cheated by optimizing the training data performed slightly better than the attack from the freshly created datasets. As the dimension increased however, the precomputed thresholds performed just as well as the attack which cheated.

When we attempt to defend against the attack, we see that calculated mean l2 norms increase with *variance on the x axis* (plots generated incorrectly). And that the attack is much less successful after we add the variance approaching the theoretical limit of 0.5 which means a random guess.

Part A

Nothing special about these plots, just expected accuracy with logistic regression.

Attack performed well to determine the proper IN and OUT sets for essentially all n. No significant difference between regularized vs unregularized.

After adding noise to the newly trained models we see that the attack accuracy significantly decreases as shown by the plots above. As the variance increases, the attack performs worse.

Part b)