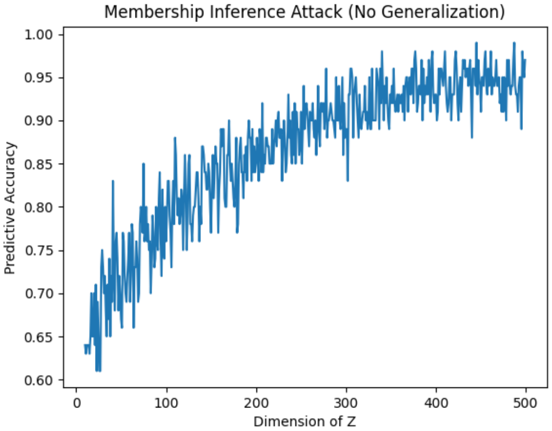
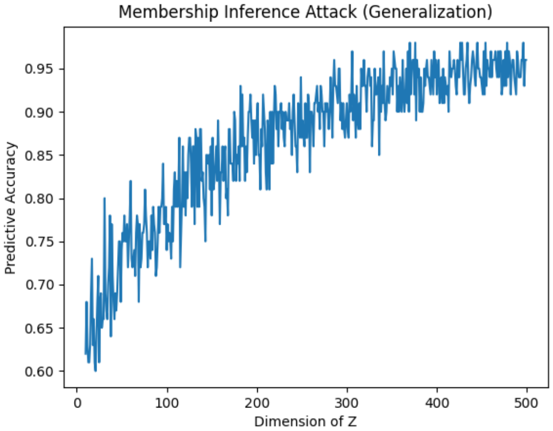
University of Waterloo
CS 480
r2knowle: 2023-11-28

Exercise # 4

Q4a) Below we can see the accuracy of the membership inference attack on both the cheating and generalized dataset:



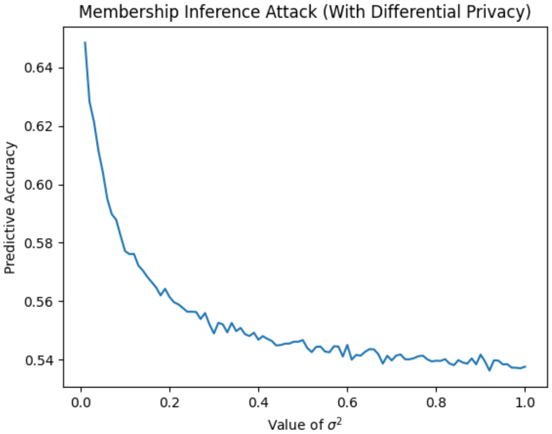
(i) Attacks on ”cheating” dataset



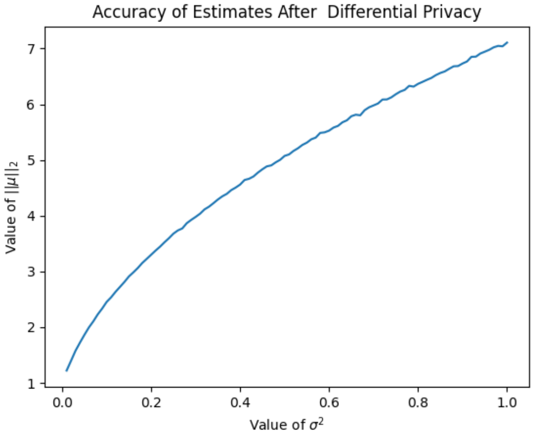
(ii) Distribution of unique testing observations

From this we can see that there appears to be an upper bound that both the cheating and generalized dataset run into, we also see that this upper bound is far better then random guessing as there would only be a 50% change of getting it right, but with this attack we can achieve accuracy’s of upto 80%.

For this next part we are going to graph the effectiveness of protecting against each attack, the first graph shows us how $||u||_2$ changes with σ which demonstrates how the accuracy of our model will change. We will also show how the effectiveness of this attack changes with different values of σ:



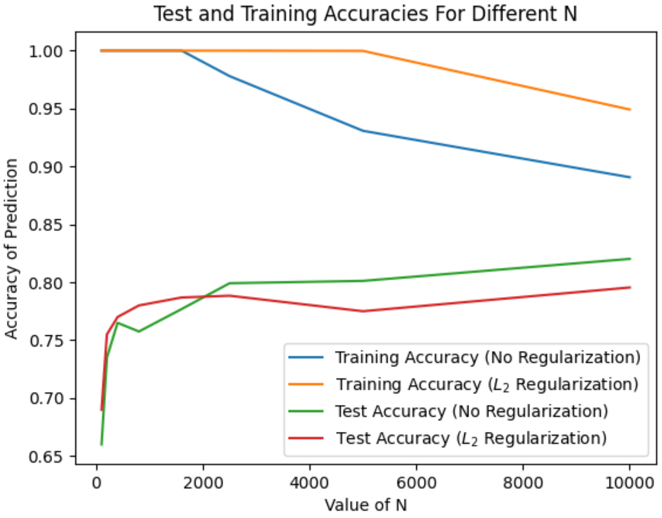
(i) Effectiveness of defences



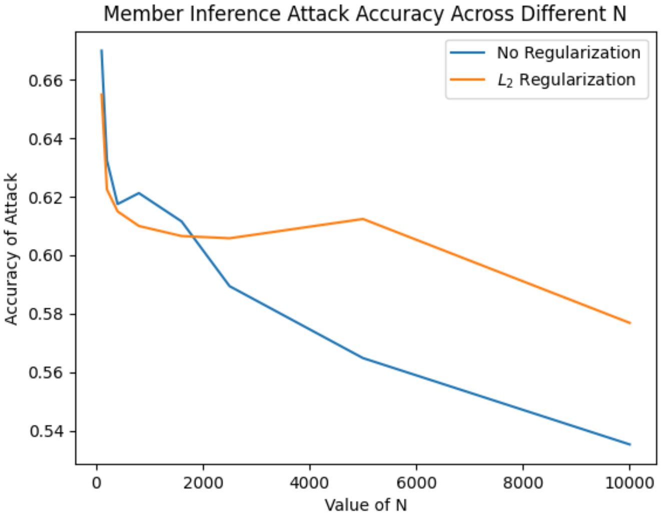
(ii) Loss in accuracy of estimates

From this we can see that as our model becomes more resistants to attacks, it also becomes less accurate. However since effectiveness of defenses has diminishing returns there is a point that maximizes both.

Q4b) Before doing any membership inference attacks, we get the following accuracies on both the test and training set using logistic regression:

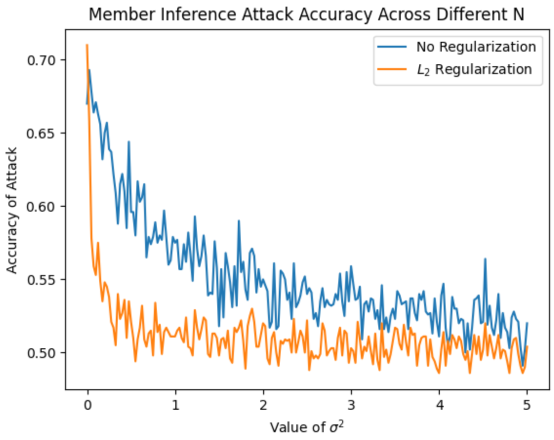


On really small N it makes sense that the training accuracy would be perfect, as each sample would be unique and so the boundary between each prediction would be large. Therefore as N gets larger it makes sense that the training accuracy would drop, it also makes sense that the test accuracy would increase and converge to a point as we are no longer under fitting.

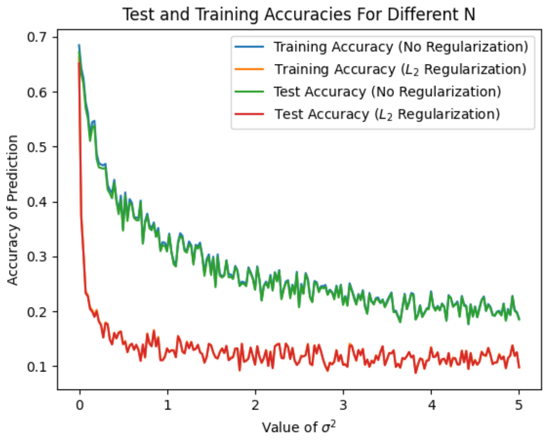


At very small Ns, as seen in the graph above our test accuracy would be very low and our training accuracy very high. This means our estimates for the IN group and OUT group would be very accurate, and so the attack accuracy would be high. As N increases and our model does worse with training and better with test it makes sense that the attack accuracy would also drop.

To defend against this attack we will be modifying the parameter vector of our logistic regression model. After doing so we get two graphs, the first shows the effectiveness of defences and the second shows the loss in accuracy of predictions:



(i) Effectiveness of defences



(ii) Loss in accuracy of estimates

We can see that predictive accuracy drops off as sigma increases, and as a result our membership inference attack becomes less and less effective. This simulates what we got from the graphs if Q4a. We notice that l2 regularization is effected more by changes in Gaussian noise then the no regularization term, and that overall because of this it is less impacted by membership inference attacks.