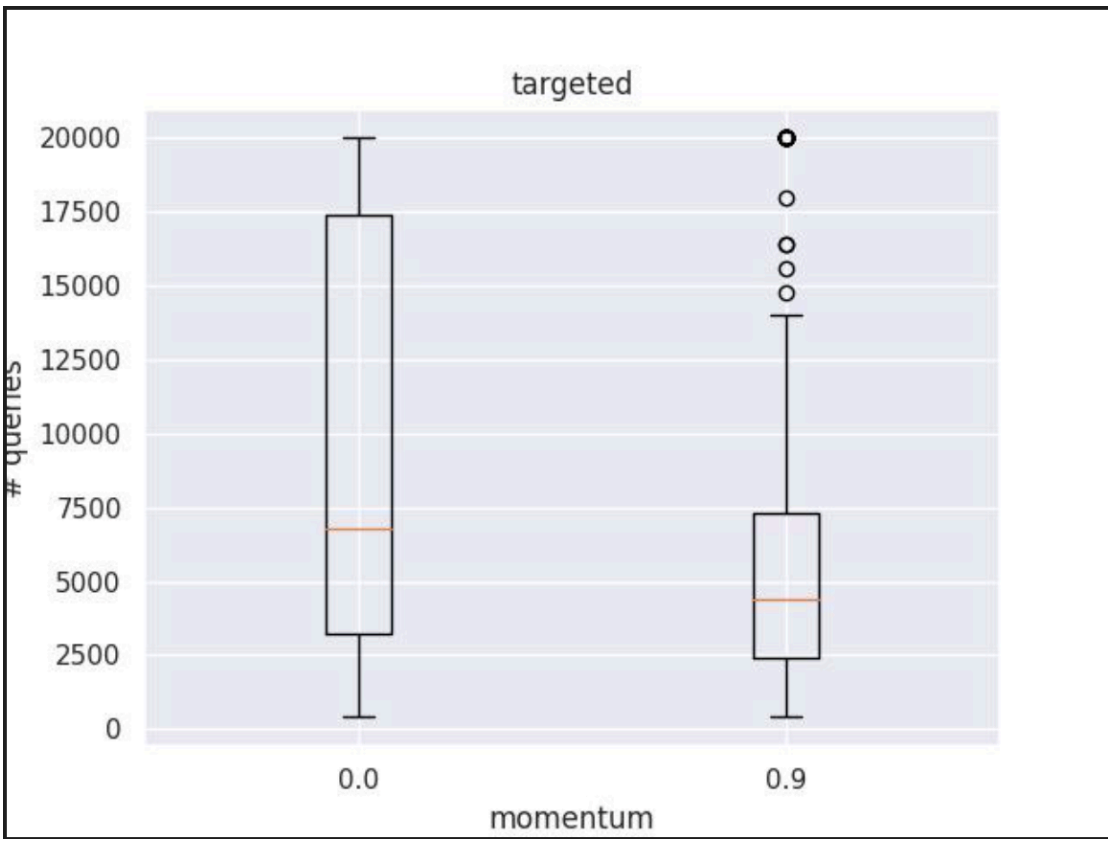
Trustworthy Machine Learning
Homework 1

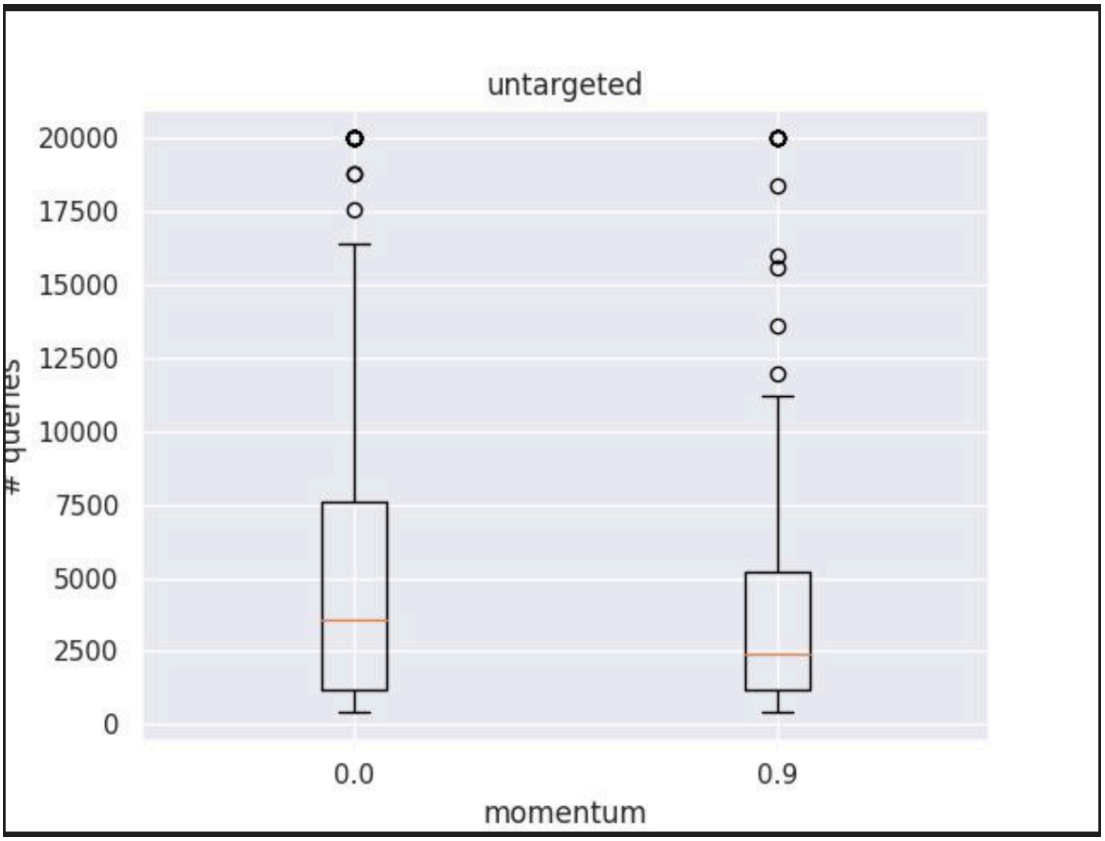
Omer Landau 206084113

(1)

1. The benign accuracy of the model is 87.50%
2. **White Box Attacks success rates:** untargeted rate is 98.50%, targeted rate is 93.50%
3. **Black Box Attacks success rates:**

	momentum=0.00	momentum=0.90
Untargeted	-success rate: 93.50% - median(# queries): 3600	-success rate: 96.50% - median(# queries): 2400
Targeted	-success rate: 76.50% - median(# queries): 6800	-success rate: 87% - median(# queries): 4400





The latter results indicates that unsurprisingly blackbox attacks are harder to apply than white box attack, in a manner that they take much longer and also has slightly lower success rate. Having said that, we are able to achieve some good results using momentum, both in the success rate and the number of queries. By incorporating momentum into the optimization process, the attacker can use the historical gradient information to smooth out the optimization trajectory, the direction is obtained by average of previous gradients, making it less sensitive to noise and random fluctuations in the input-output pairs.

(2)

1. **Test Accuracy:**
    1. Model 0: 87.50%
    2. Model 1: 82.50%
    3. Model 2: 79%
2.

Untargeted	Model 0	Model 1	Model 2
Model 0	0.985	0.57	0.54
Model 1	0.69	0.965	0.585
Model 2	0.59	0.545	0.955

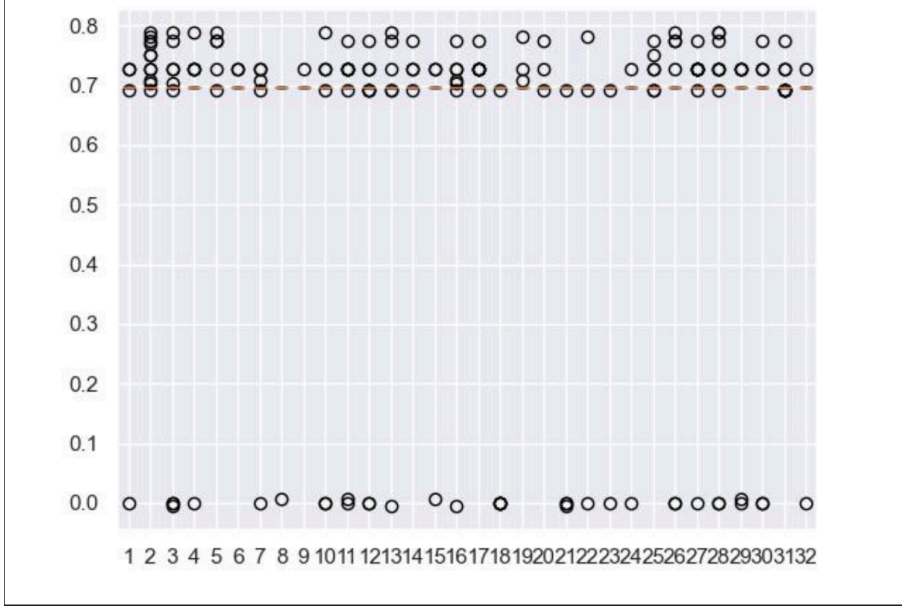
Targeted	Model 0	Model 1	Model 2
Model 0	0.96	0.295	0.275
Model 1	0.39	0.9	0.275
Model 2	0.355	0.25	0.86

3. Ensemble attacks' transferability from models 1+2 to model 0:
   Untargeted attack: 0.7400
   Targeted attack: 0.4900

Using ensemble attacks improved the transferability of both targeted and untargeted attacks. In the case of untargeted attacks, the success rate when transferring from model 1 and 2 to 0 increased from 0.69 and 0.59 respectively to 0.74 when using from their ensemble. Similar improvements can be seen for targeted attacks. This overall success might be due to the fact that the adversarial attack is compelled to discover more general perturbations that do not depend on the specifics of one model, hence by using an ensemble we are considering the expected losses and unique specification of couple of models.

(3)

1. Maximal RAD: 0.6970 (on GPU) and 0.7879 (on CPU)
2. 2.60% of bits flipped are leading to >15% RAD.
3. We can see that there is impact on both LSBits and MSBits, though the 2nd bit has probably the highest RAD. This makes total sense, In a 32-bit floating point number, the first bit represents the sign, the next 8 bits represent the exponent, and the remaining 23 bits represent the mantissa or fraction. Flipping the second bit of the exponent from 1 to 0 reduces the exponent by 128, which corresponds to a division by 2^128. Also, since a lot of the weights are relatively small, this bit would be 0 most of the times. Hence, flipping it to 1 will make the weight much larger than flipping any other bits, it will explode the weight.


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32