Trustworthy Machine Learning, Spring 2023
Homework Assignment 1
Eitan Levinzon, 209044841

Question 1

Section 1

The benign accuracy of the model is 0.8750.

Section 2

The success rate of the untargeted white-box attack is 0.9850.
The success rate of the targeted white-box attack is 0.9450.

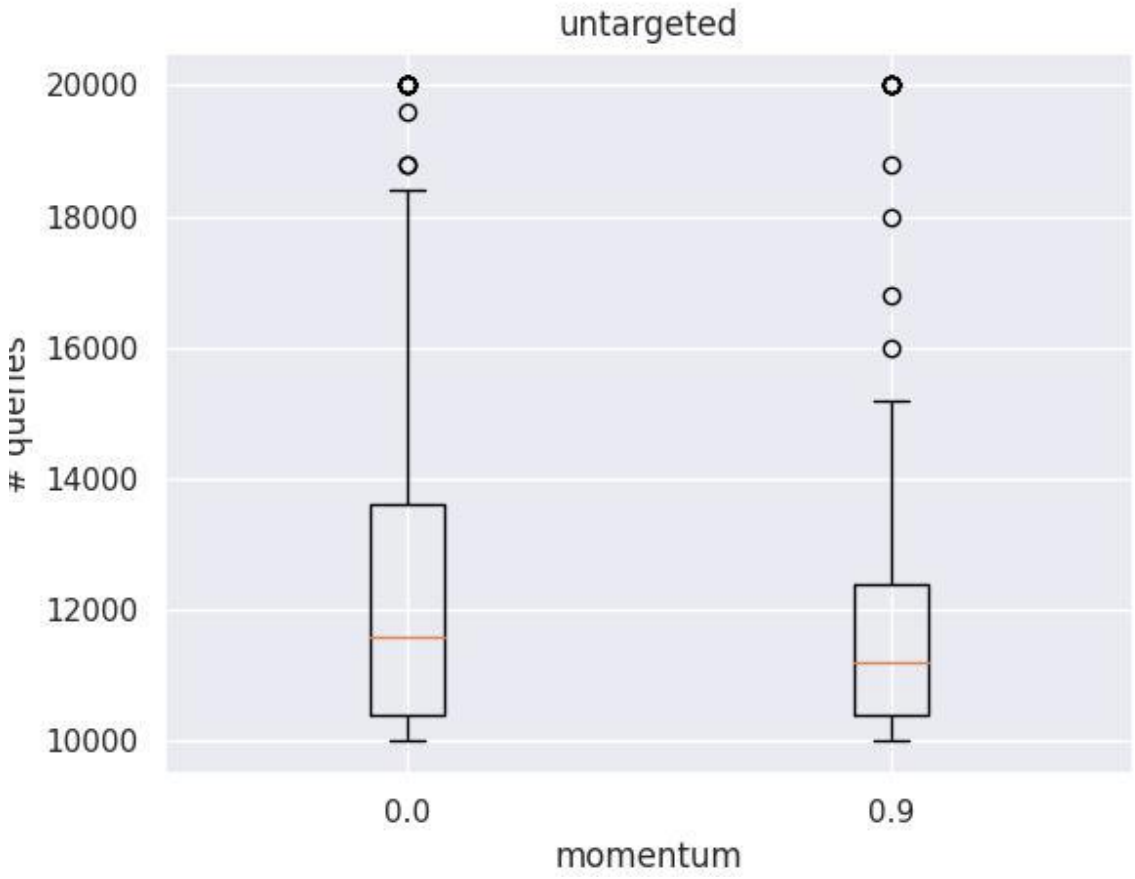
Section 3

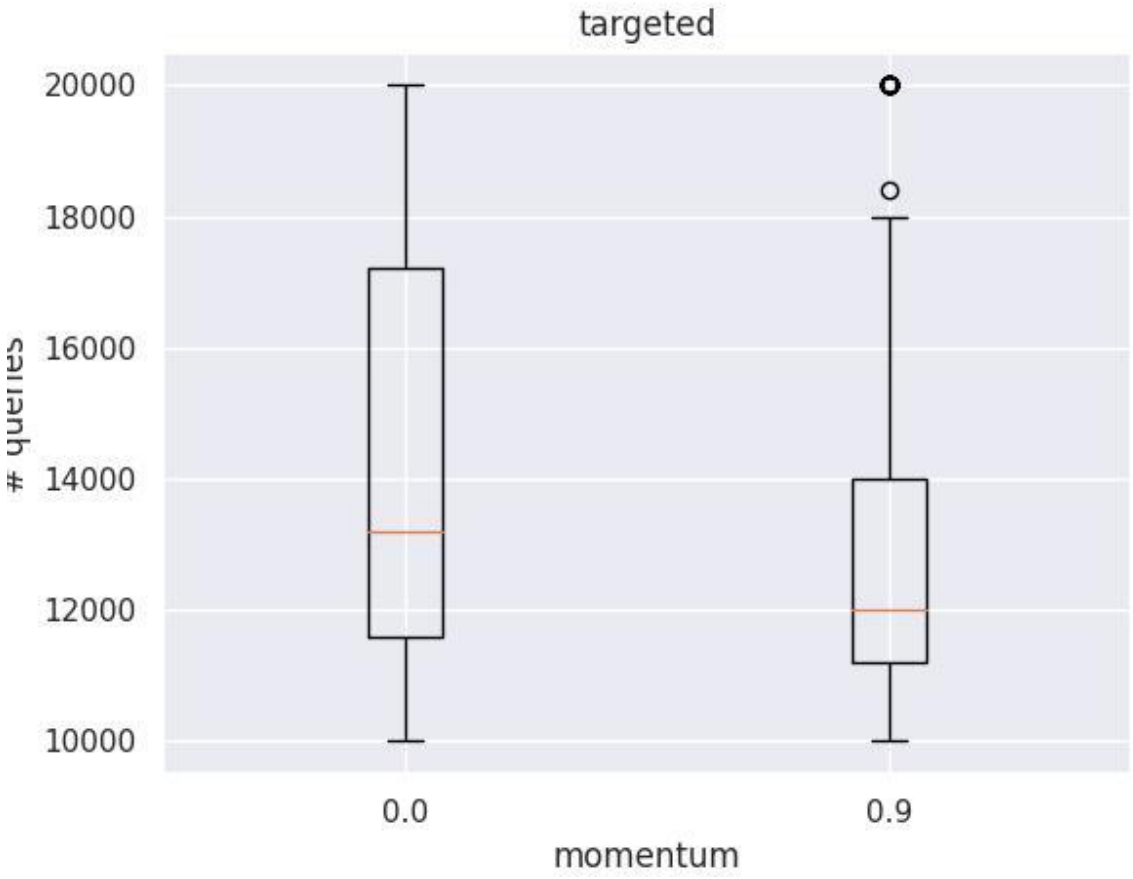
Success rates of the black-box attack:

- Without momentum
    - Untargeted: 0.9400
    - Targeted: 0.8000
- With momentum
    - Untargeted: 0.9650
    - Targeted: 0.8700

The success rates are, as expected, worse for the black-box attack compared to the white-box one.
We can see that momentum improves the results, though they're still less than the white-box ones. This is probably due to the fact that momentum helps incorporate many queries together, instead of giving the most effect to the latest queries, thus creating a more stable and accurate gradient average.

See plots:


untargeted
20000
18000
16000
14000
12000
10000
queries
0.0
0.9
momentum



Question 2

The results of transferring untargeted attacks are:

	On Model 1	On Model 2	On Model 3
From Model 1	0.985	0.57	0.535
From Model 2	0.685	0.965	0.59
From Model 3	0.59	0.545	0.95

And the results of transferring targeted attacks are:

	On Model 1	On Model 2	On Model 3
From Model 1	0.92	0.255	0.26
From Model 2	0.38	0.895	0.275

From Model 3	0.325	0.235	0.84

From the results we conclude several things:

- Transferability is much better on untargeted attacks. This is probably due to the fact that making a model err in any way is much easier than making it err in a specific way, and the specific way's method would probably not transfer very well to other models, hence the low success rate on targeted attacks.
- However, targeted attacks do transfer. Specifically, we can see that model 1 was most susceptible to this kind of attack.

Ensemble attacks' transferability's success rate:

- Untargeted: 0.7450
- Targeted: 0.4750

We can see that the results did improve using the ensemble method, both in the targeted and the untargeted cases.
In the ensemble attack we attacked model 1, and the results improved from ~0.63 to 0.74 in the untargeted case and from ~0.35 to 0.47.
In both cases we saw a similar increase of ~0.13.
We saw that model 1 was most susceptible to transferability-based attacks, so it's not surprise that using an ensemble model that improves from several models is successful when attacking model 0.
Using an ensemble model helps us find adversarial examples that are more general, and thus transfer better.

Question 3

Section 1

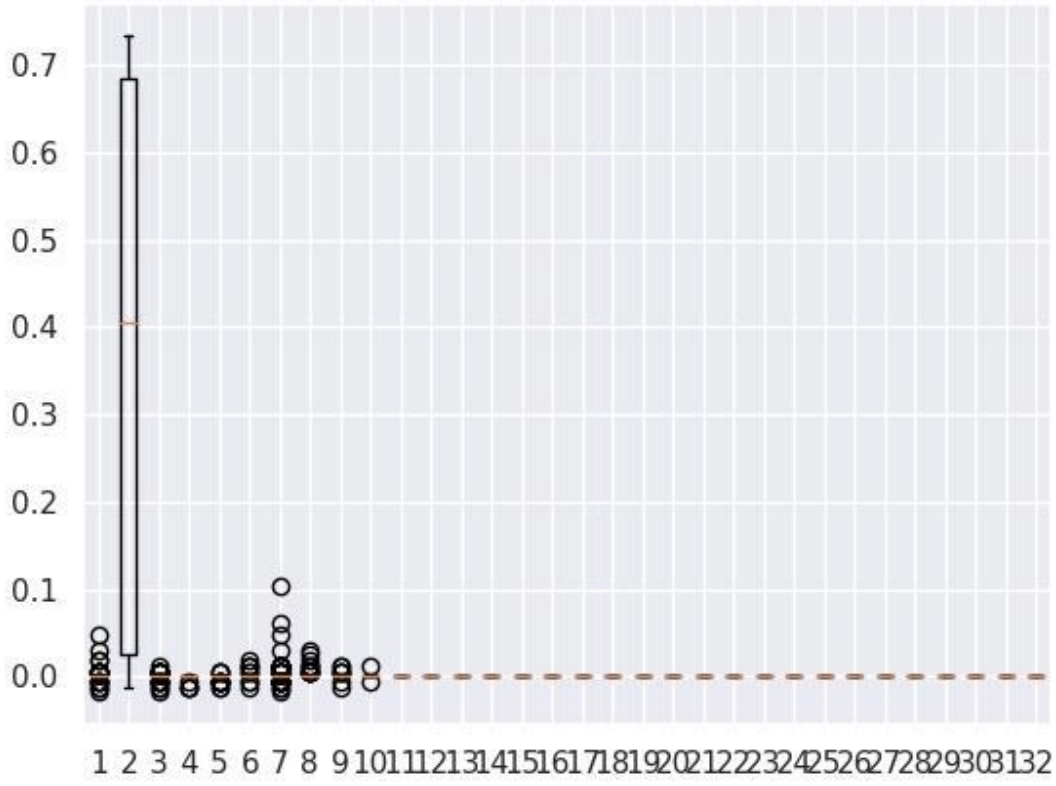
The maximum RAD achieved is 0.7333.

Section 2

Only 2.36% of the bits flipped lead to >15% RAD.

Section 3

The second bit has the highest effect and median RAD. This is due to the representation of float numbers in 32 bits values - the second bit is the highest value of the exponent of the value of the number. Changing changes the result the most.
See plot:


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