Yoav Gur Arieh | 318637592

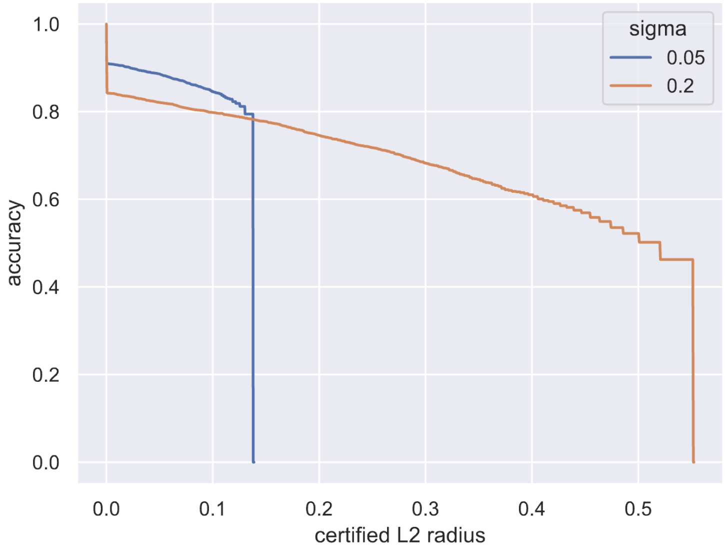
Question 1

3. It took 1390 seconds to train the standard model, and 1697 seconds to train the free adversarial model.
4. The Accuracy of the standard model is 0.9185, and the accuracy of the free adversarial model is 0.8972. The sucess rate of PGD of the standard model is 0.9008 and of the free adversarial model is 0.2865. Thus we can see that adversarial training hurts benign accuracy but makes the model much more robust to attacks.
5. We can see that incresing *m* increases training time and robustness, while generally having positive effects on accuracy (the trend is positive).

m	Accuracy	Success Rate of PGD	Training Time (secs)
4	0.8972	0.2865	1697
5	0.8975	0.2753	1860
6	0.8970	0.2720	2139
7	0.8995	0.2703	2372

Question 2

4. We can see that at sigma=0.05 we have a higher certified robustness at first, until reaching a radius of approximately 0.14. At that point sigma=0.2 overtakes in terms of accuracy and maintains it until approximately 0.57. This is as expected - the higher sigma value leads to more significant noise, making the model generally more robust to larger perturbations, while sacrificing accuracy.

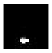


Question 3

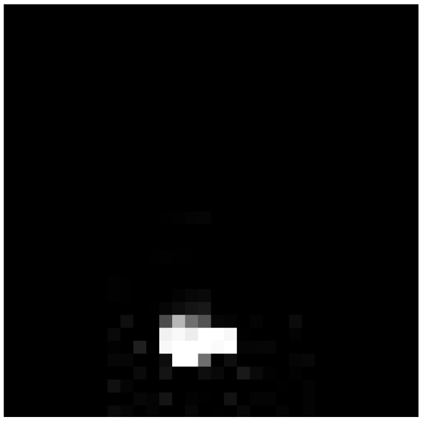
2. We can see that both outputs have a similar accuracy, indicating that their performance on clean data is nearly identical, meaning we'd be unlikely to be able to distinguish between them based on this metric. We can also clearly identify that the class and model for which a much smaller trigger is needed to misclassify samples into this target label are model 1 and class `0` (with a norm of 38). We can also see that the backdoor is highly effective, succeeding in all tests.

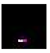
```
yoavgurarieh@c-001:~/hw2$ python main_c.py
Accuracy of model 0: 0.9168
Accuracy of model 1: 0.9107
Norm of trigger targeting class 1 in model 0: 106.2822
Norm of trigger targeting class 2 in model 0: 148.8787
Norm of trigger targeting class 3 in model 0: 135.0353
Norm of trigger targeting class 0 in model 1: 38.5321
Norm of trigger targeting class 1 in model 1: 123.3861
Norm of trigger targeting class 2 in model 1: 156.0660
Norm of trigger targeting class 3 in model 1: 145.0165
Which model is backdoored (0/1)? 1
Which class is the backdoor targeting (0/1/2/3)? 0
Backdoor success rate: 1.0000
```

3. Answer and images included below:
   1. We can see that the backdoor consists of a small rectangle at the bottom of the image, with a purplish hue to part of it. This can be seen from the trigger as well as the mask which restricts to that same part of the image.
   2. As previously noted, the model with the backdoor has an accuracy that, while slightly lower than the normal model, is very similar to it. This difference in accuracy is small enough that the backdoor can be considered as highly effective, as the difference is so small.
   3. We can also see that the backdoor success rate, i.e. its success at causing misclassification as the target class, is extremely effective, standing at 1.0.

   Mask: 

   Mask zoomed in to be able to see better:



Trigger:

Trigger zoomed in to be able to see better: