TML HW2

Omer Landau
206084113
omerlandau1@mail.tau.ac.il

Q1 Free Adversarial Training

3

```
Training standard model...
Time (in seconds) to complete standard training: 453.5353
Adversarially training a model...
Time (in seconds) to complete free adversarial training: 461.2511
```

4

```
Model accuracy:
    - standard    : 0.9187
    - adv_trained : 0.8127
Success rate of untargeted white-box PGD:
    - standard  : 0.8950
    - adv_trained: 0.3838
```

The adversarial training time compared to standard training is pretty much the same, which is coherent with the paper's results and purpose ("free" adversarial training). We can see that the benign accuracy of the adversarial model is inferior to the standard model by approx 10%. Having said that, it is 50% less vulnerable to PGD attacks, hence more robust.

5

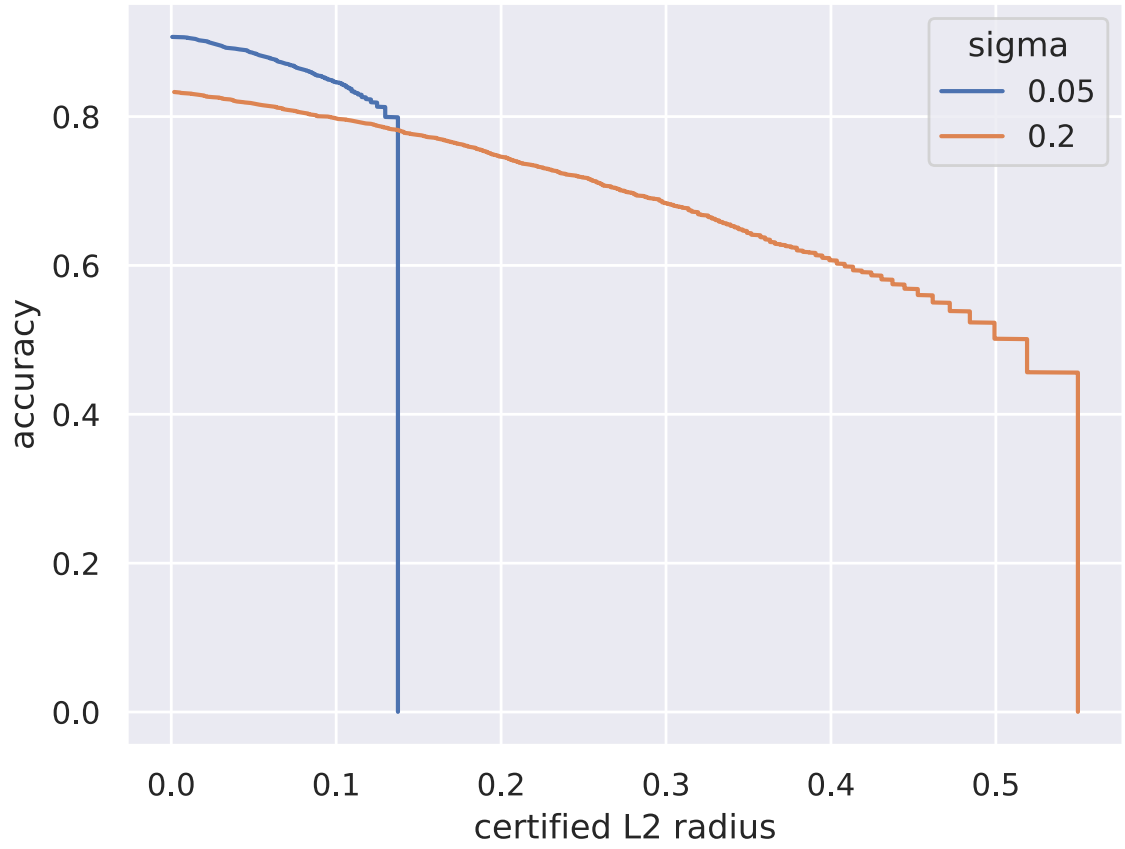
```
Adversarially training a model...
Time (in seconds) to complete free adversarial training: 443.4370

Model accuracy:
    - standard    : 0.9187
    - adv_trained : 0.7640
Success rate of untargeted white-box PGD:
    - standard  : 0.8950
    - adv_trained: 0.4080
```

Increasing the m parameter from 4 to 7 improved training time and it is faster by a few percents, on the other hand it decreased both accuracy and robustness

as the benign accuracy is inferior by almost 5%. And the success rate of PGD against the model increased by 2%.

Q2 Randomized Smoothing

4



From the experiment we can see, training with larger sigma can certify accuracy for perturbations with large L2 radii. While the main tradeoff is the accuracy in overall, which is lower than model trained with smaller sigma (even on small radii). Additional conclusion is that for both sigma values, it is more difficult to the certify samples when the L2 radii is bigger.

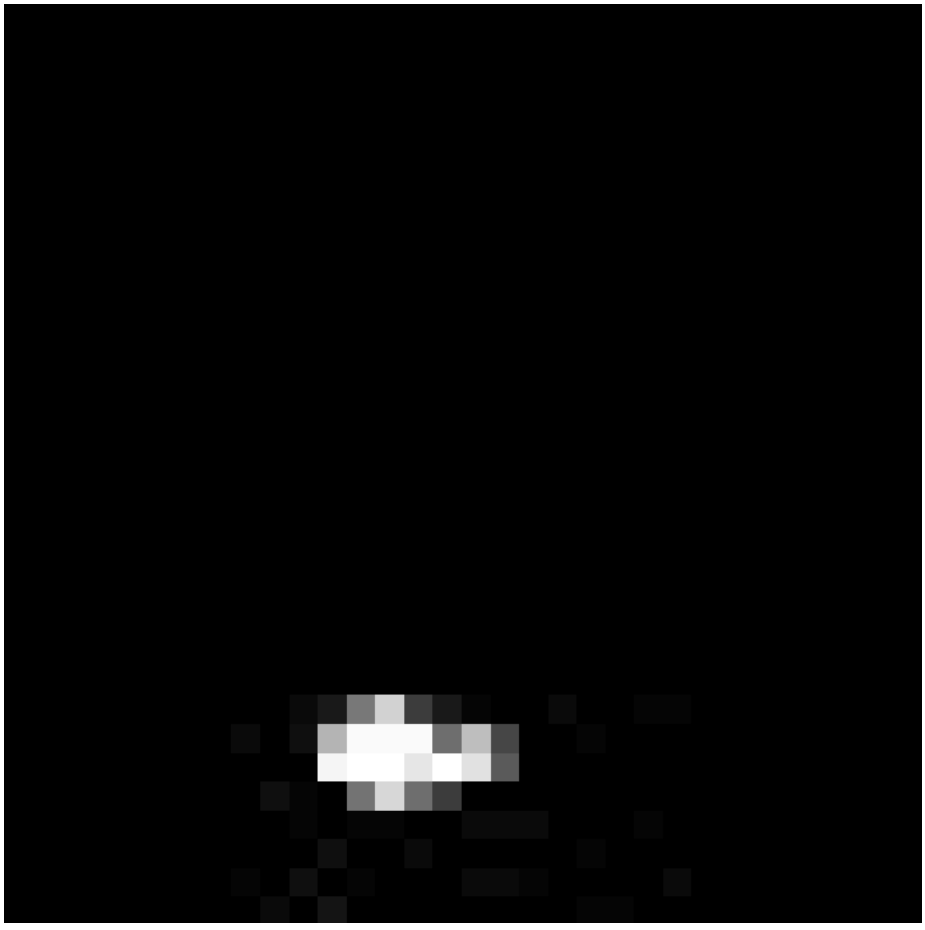
Q3

2

```
Accuracy of model 0: 0.9168
Accuracy of model 1: 0.9107
Norm of trigger targeting class 0 in model 0: 176.5310
Norm of trigger targeting class 1 in model 0: 137.7255
Norm of trigger targeting class 2 in model 0: 207.9565
Norm of trigger targeting class 3 in model 0: 189.3265
Norm of trigger targeting class 0 in model 1: 52.4745
Norm of trigger targeting class 1 in model 1: 175.3760
Norm of trigger targeting class 2 in model 1: 206.0672
Norm of trigger targeting class 3 in model 1: 191.4642
Which model is backdoored (0/1)? 1
Which class is the backdoor targeting (0/1/2/3)? 0
Backdoor success rate: 0.9982
```

I chose the model and class using the norm, we optimize on the size of the trigger, hence we believe that the trigger of the backdoored model should be small.

Mask



Backdoor

3

1

It is clear that the backdoor is small as wished, being a small purple rectangle located in the lower middle part of the image. It can be hard to observe when added to a benign image.

2

Undoubtedly, the benign accuracy of the backdoored model closely resembles that of the non-backdoored model. The disparity in accuracy between the models can be attributed to both random factors and a small influence of the backdoor.

The backdoor proves to be highly effective in causing misclassification, achieving an impressively high success rate of over 99.8%.