

1 Question 1

1.3

Time (in seconds) to complete standard training: 1359.3467.

Time (in seconds) to complete free adversarial training: 372.3792.

1.4

| Model/Task | Accuracy | PGD Success rate |
|--------------|----------|------------------|
| Standard | 0.9170 | 0.8928 |
| Adv.-trained | 0.8087 | 0.3763 |

Adversarial training significantly decreases the PGD success rate, i.e. increases robustness with the price of a small decrease in accuracy.

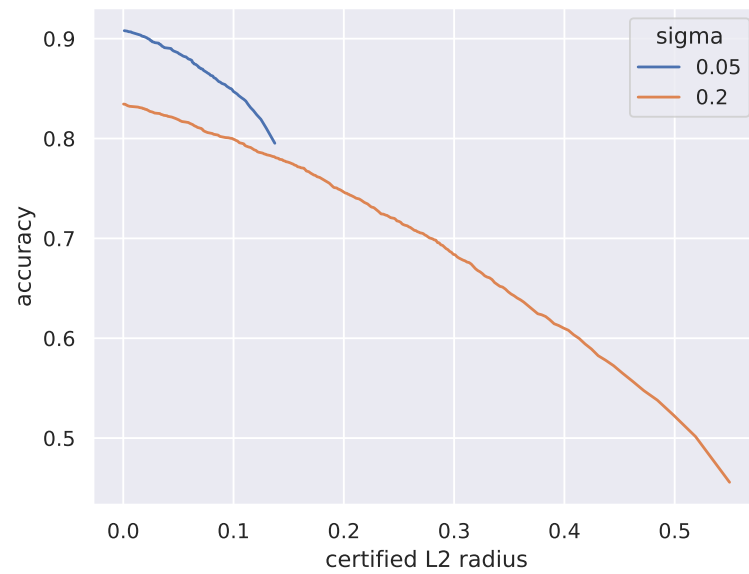
1.5

| m/metric | Time | Accuracy | PGD Success rate |
|----------|----------|----------|------------------|
| 4 | 426.3014 | 0.8083 | 0.3790 |
| 5 | 351.1917 | 0.7847 | 0.3868 |
| 6 | 290.3951 | 0.7682 | 0.3953 |
| 7 | 268.9399 | 0.7483 | 0.4095 |

It seems that both the benign accuracy, robustness and training time decrease as m increases.

2 Question 2

2.2



3 Question 3

3.2

Model 1 is backdoored, the backdoor forces it to output class 0.

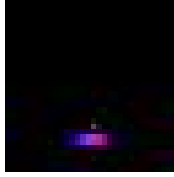
3.3

3.3.1

mask:



trigger:



3.3.2

Yes. The accuracy of Model 1 is only very slightly lower than that of Model 2 (0.9107 vs 0.9168).

3.3.3

Very successful, its success rate is 0.9978.