

1 Question 1

1.3

Time (in seconds) to complete standard training: 1381.1138.

Time (in seconds) to complete free adversarial training: 386.9370.

1.4

Model/Task	Accuracy	PGD Success rate
Standard	0.9168	0.9025
Adv.-trained	0.8225	0.3695

Adversarial training significantly decreases the PGD success rate, i.e. increases robustness with the price of a some (around 10%) decrease in accuracy.

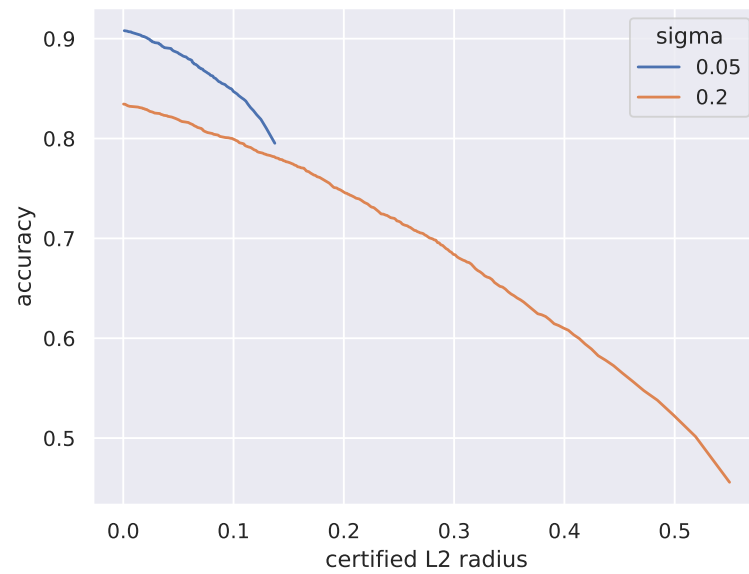
1.5

m/metric	Time	Accuracy	PGD Success rate
4	383.8422	0.8077	0.3808
5	323.3860	0.7900	0.3828
6	276.1508	0.7713	0.3963
7	258.2459	0.7460	0.4073

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

2 Question 2

2.2



3 Question 3

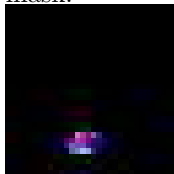
3.2

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

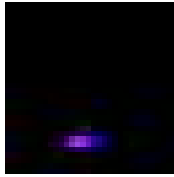
3.3

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 0.9982.