A Sample Complexity Rates for Adversarial Training

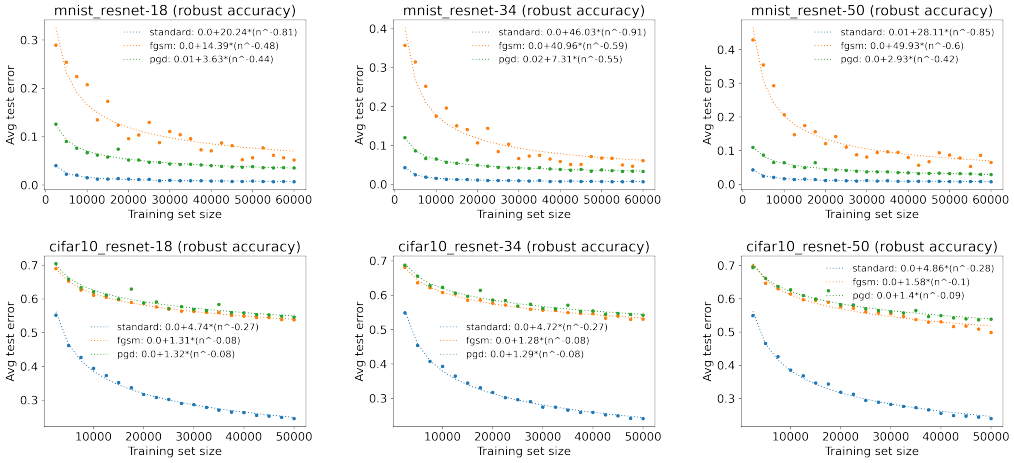

Figure 3: Sample complexity rates of varying-depth ResNets trained on MNIST and CIFAR-10 datasets, including both standard training and adversarial training.

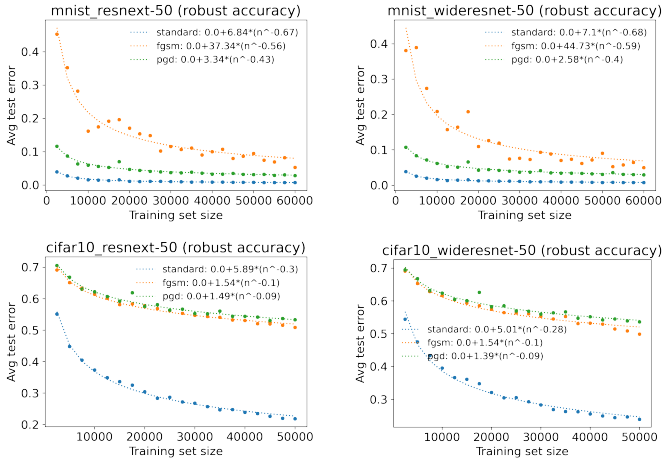

Figure 4: Sample complexity rates of ResNeXt-50 32x4d and Wide ResNet-50 2x trained on MNIST and CIFAR-10 datasets, including both standard training and adversarial training.

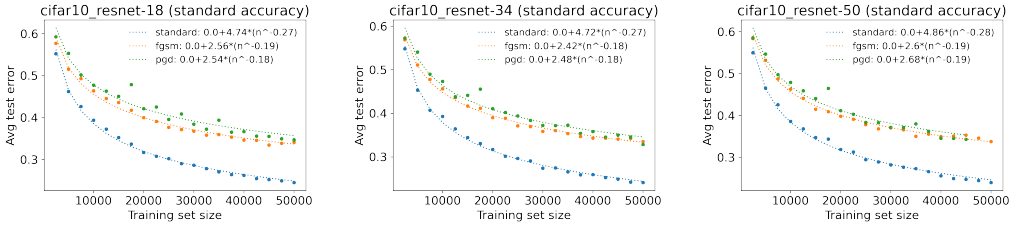

Figure 5: Sample complexity rates of varying-width ResNets trained on CIFAR-10 datasets, including both standard training and adversarial training. "Standard accuracy" indicates that no attacks are used at test time.

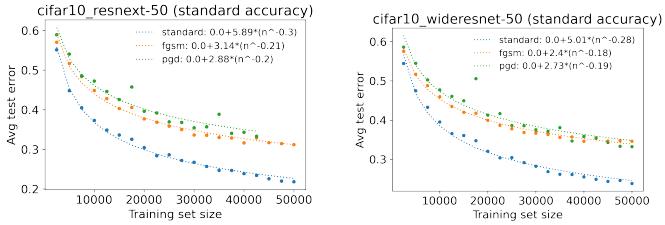

Figure 6: Sample complexity rates of ResNeXt and Wide ResNet trained on CIFAR-10 datasets, including both standard training and adversarial training. "Standard accuracy" indicates that no attacks are used at test time.

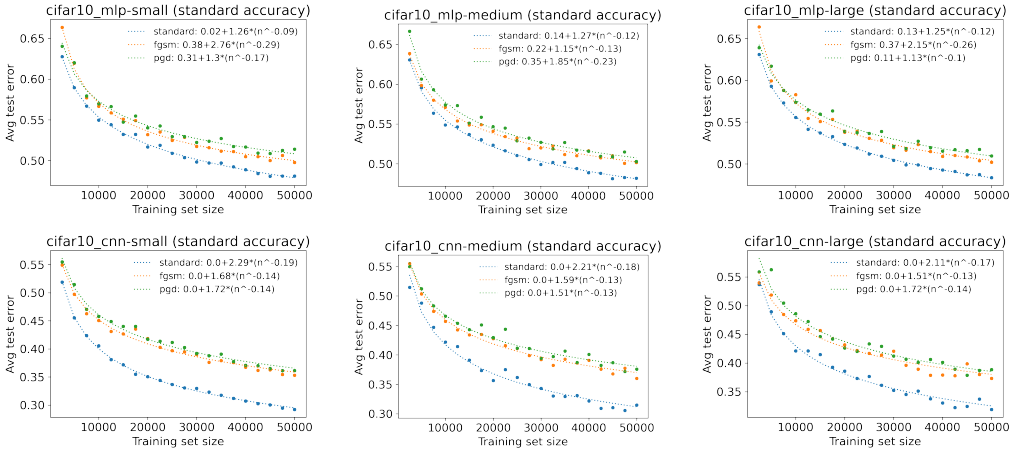

Figure 7: Sample complexity rates of varying-width two-layer MLPs and simple CNNs trained on CIFAR-10 datasets, including both standard training and adversarial training. "Standard accuracy" indicates that no attacks are used at test time.

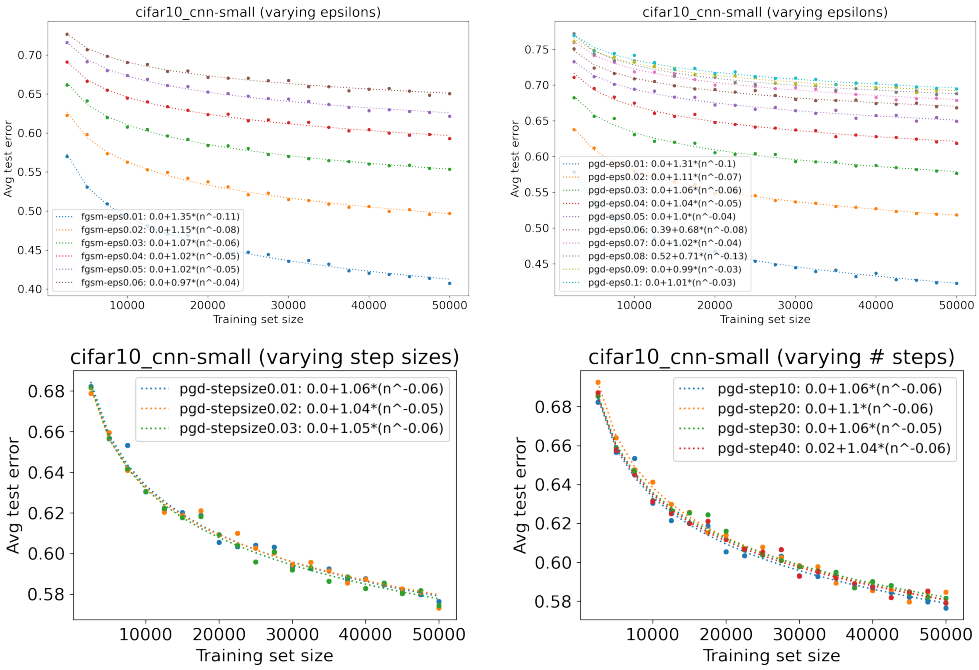

Figure 8: Sample complexity rates when tuning ϵ for FGSM and ϵ, step size, and the number of steps for PGD. The upper-left figure for FGSM only has ϵ up to 0.07, because values beyond that produce noisy curves that are infeasible to fit. Because of the computation constraint, this is done on the CNN-Small model.

B Number of Data Points Required for Adversarial Training

Model	Standard	FGSM	PGD
MLP-Small	100	8	603
MLP-Medium	100	8	1668
MLP-Large	100	14	3981
CNN-Small	100	39	1557
CNN-Medium	100	36	1166
CNN-Large	100	1101	954
ResNet-18	100	22	4806
ResNet-34	100	2371	2037
ResNet-50	100	681	11159
ResNeXt-50-32x4d	100	247	1307
Wide ResNet-50-2	100	202	2512

Table 1: Number of training examples required for adversarial training to reach the same robust accuracy as standard training on 100 examples using MNIST.

Model	Standard	FGSM	PGD
MLP-Small	100	63	1.00e+03
MLP-Medium	100	1.00e+04	1.00e+03
MLP-Large	100	7	52
CNN-Small	100	2.20e+07	2.20e+07
CNN-Medium	100	1.00e+06	3.16e+04
CNN-Large	100	7.20e+04	4.64e+05
ResNet-18	100	5.60e+06	5.60e+06
ResNet-34	100	5.60e+06	5.60e+06
ResNet-50	100	3.98e+05	1.67e+06
ResNeXt-50-32x4d	100	1.00e+06	4.64e+06
Wide ResNet-50-2	100	3.98e+05	1.67e+06

Table 2: Number of training examples required for adversarial training to reach the same robust accuracy as standard training on 100 examples using CIFAR-10.

Model	Standard	FGSM	PGD
MLP-Small	100	4	11
MLP-Medium	100	70	11
MLP-Large	100	8	251
CNN-Small	100	518	518
CNN-Medium	100	588	588
CNN-Large	100	412	412
ResNet-18	100	695	1000
ResNet-34	100	1000	1000
ResNet-50	100	886	886
ResNeXt-50-32x4d	100	720	1000
Wide ResNet-50-2	100	1292	886

Table 3: Number of training examples required for adversarial training to reach the same standard accuracy as standard training on 100 examples using CIFAR-10.

C Effects of Epsilon, Step Size, and Number of Steps

Epsilon	Standard	FGSM	PGD
0.01	-0.19	-0.11	-0.1
0.02	-0.19	-0.08	-0.07
0.03	-0.19	-0.06	-0.06
0.04	-0.19	-0.05	-0.05
0.05	-0.19	-0.05	-0.04
0.06	-0.19	-0.04	-0.08
0.07	-0.19	n/a	-0.04
0.08	-0.19	n/a	-0.13
0.09	-0.19	n/a	-0.03
0.1	-0.19	n/a	-0.03

Table 4: Sample complexity rates in adversarial training and standard training with CNN-Small as ϵ increases. Some values of the FGSM column are n/a because the curves are too noisy to fit.

# Steps	Standard	PGD
10	-0.19	-0.06
20	-0.19	-0.06
30	-0.19	-0.05
40	-0.19	-0.06

Table 5: Sample complexity rates in adversarial training and standard training with CNN-Small as the number of steps increases.

Step Size	Standard	PGD
0.01	-0.19	-0.06
0.02	-0.19	-0.05
0.03	-0.19	-0.06

Table 6: Sample complexity rates in adversarial training and standard training with CNN-Small as the step size increases.