Sample Complexity Scaling Laws for Adversarial Training

Chris Liu

December 1, 2021

Outline

- Recap on background
- Experiments
- Results
- Conclusion

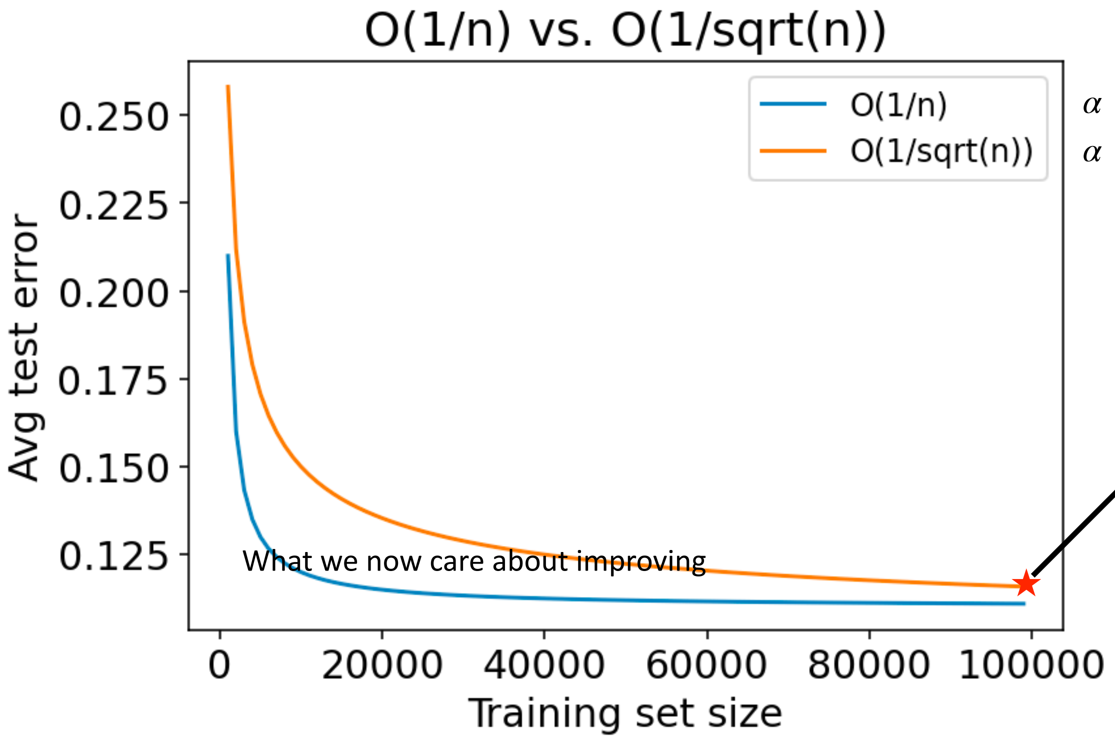
(Empirical) Sample Complexity Rate

- Given a model f_θ trained with n examples, it can achieve ϵ generalization error, where ϵ is characterized by a power-law function $\mathscr{E}$ of training data size n.

$$\mathscr{E}_{f_\theta}(n) = \alpha + \beta \cdot n^\gamma \in O(n^\gamma)$$

- α, β, γ are constants.
  - $O(\sqrt{n})$: γ remains approximately $\mathbf{-1/2}$ in most real-world agnostic settings.
  - $O(1/n)$: γ can reach as fast as $\mathbf{-1}$ (or faster) with additional assumptions (e.g., Tsybakov low noise condition, (Tsybakov, 2004)).
  - The rate is asymptotic (i.e., $\forall n \geq n_0$).

(Empirical) Sample Complexity Rate



$\alpha = 0.11,\ \beta = 100,\ \gamma = -1.0$

$\alpha = 0.1,\ \beta = 5,\ \gamma = -0.5$

What we typically care about improving

For the orange curve to achieve an expected error rate of 12.5%, it needs 40000 training samples, whereas the blue curve only requires 6667 (83.3% less) training samples.

Objectives

1. *Does adversarial training affect the scaling law of sample size vs. robust error?*

2. *Does adversarial training affect the scaling law of sample size vs. standard error?*

3. *If adversarial training reduces sample efficiency, what is the cause?*

4. *How much more data points do we need for adversarial training to reach the same robust/standard error as standard training?*

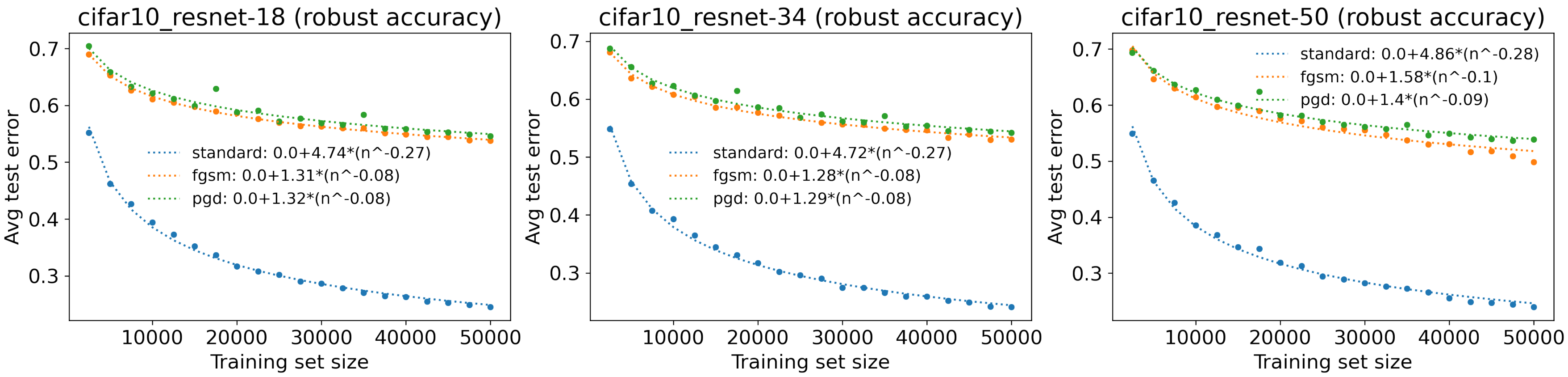
Experiment Setting

- Datasets: MNIST and CIFAR-10
- Adversarial training: FGSM and PGD
  - $\epsilon = 0.3/0.03,\ \alpha = 0.01,\ 10$ steps (by default)
- Models
  - MLP (2-layer), simple CNN (2-layer)
    - “Medium” and “Large” MLPs and CNNs have 2x and 4x widths.
  - ResNet-18/34/50, ResNeXt-50-32x4d, Wide ResNet-50x2
- No other training tricks are applied except early stopping.

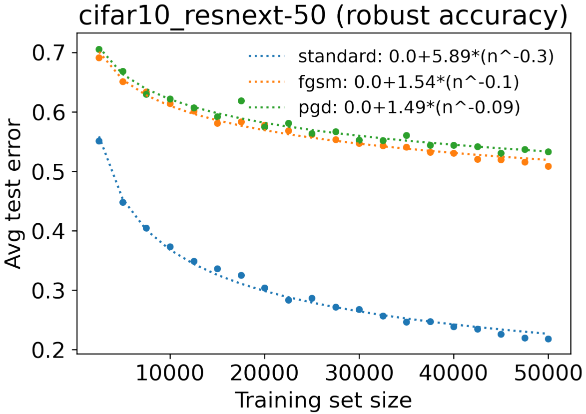
Experiment Setting

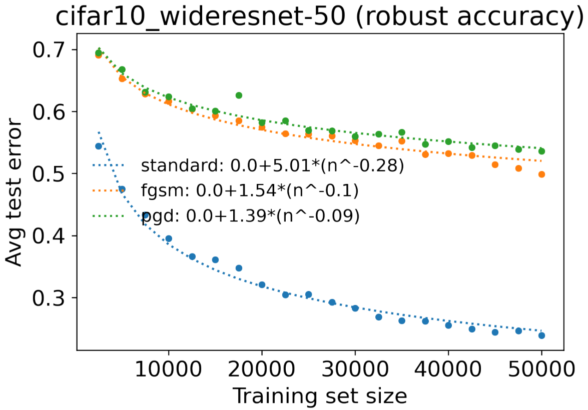
- Measuring the empirical sample complexity rate
    1. Sample a subset of the original training set
        1. 2500, 5000, …
    2. Train a randomly initialized model on the subset with early stopping for 10 trials
    3. Calculate average error of 10 models
    4. Fit the resulting curve

CIFAR-10 (Robust Accuracy)



CIFAR-10 (Robust Accuracy)

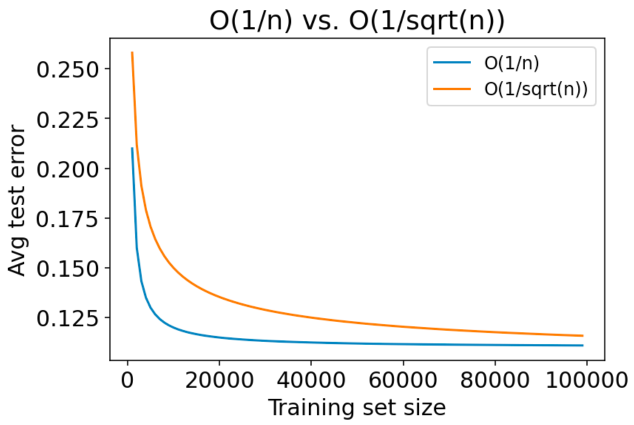




(Empirical) Sample Complexity Rate

- Given a model f_θ trained with n examples, it can achieve ϵ generalization error, where ϵ is characterized by a power-law function $\mathscr{E}$ of training data size n.

$$\boxed{\mathscr{E}_{f_\theta}(n) = \alpha + \beta \cdot n^\gamma} \in O(n^\gamma)$$



CIFAR-10 (Robust Accuracy)

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

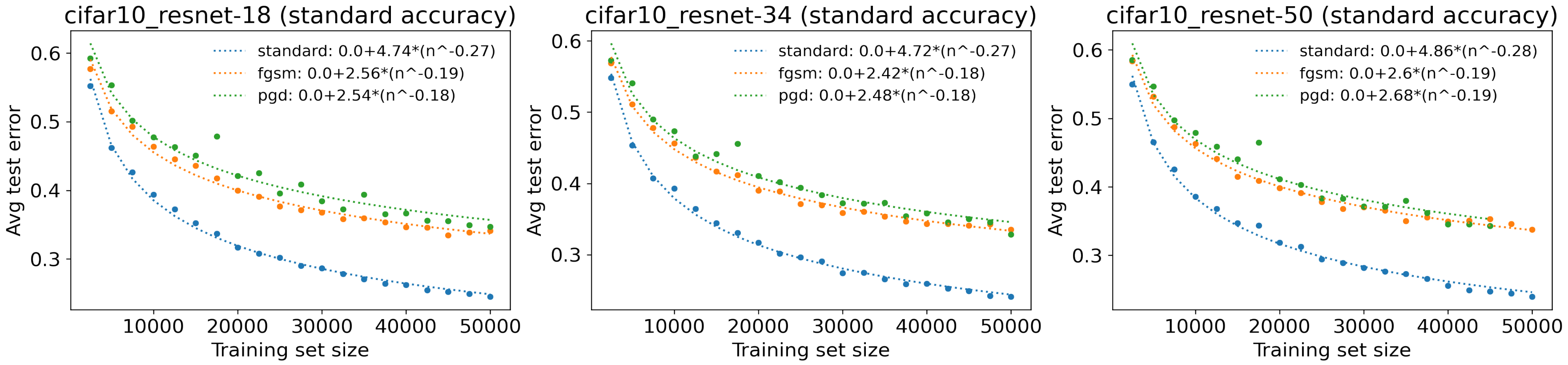
Table 2: Number of training examples required for adversarial training to reach the same robust accuracy as standard training on 100 examples using CIFAR-10.

MNIST (Robust Accuracy)

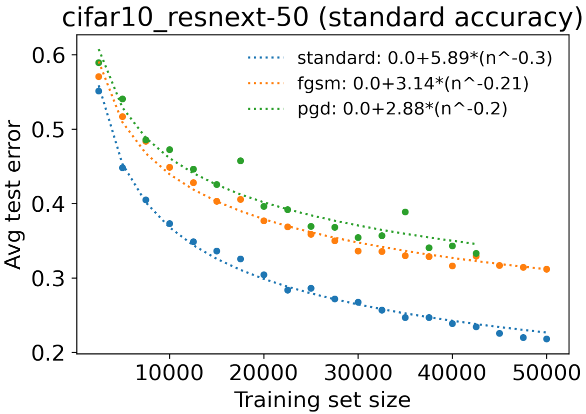
Model	Standard	FGSM	PGD
MLP-Small	100	8	603
MLP-Medium	100	8	1668
MLP-Large	100	14	3981
CNN-Small	100	39	1557
CNN-Medium	100	36	1166
CNN-Large	100	1101	954
ResNet-18	100	22	4806
ResNet-34	100	2371	2037
ResNet-50	100	681	11159
ResNeXt-50-32x4d	100	247	1307
Wide ResNet-50-2	100	202	2512

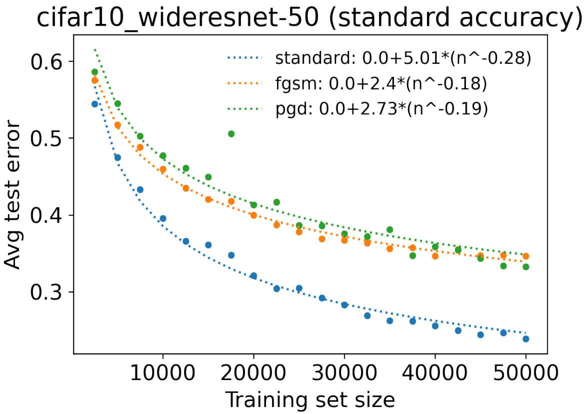
Table 1: Number of training examples required for adversarial training to reach the same robust accuracy as standard training on 100 examples using MNIST.

CIFAR-10 (Standard Accuracy)



CIFAR-10 (Standard Accuracy)



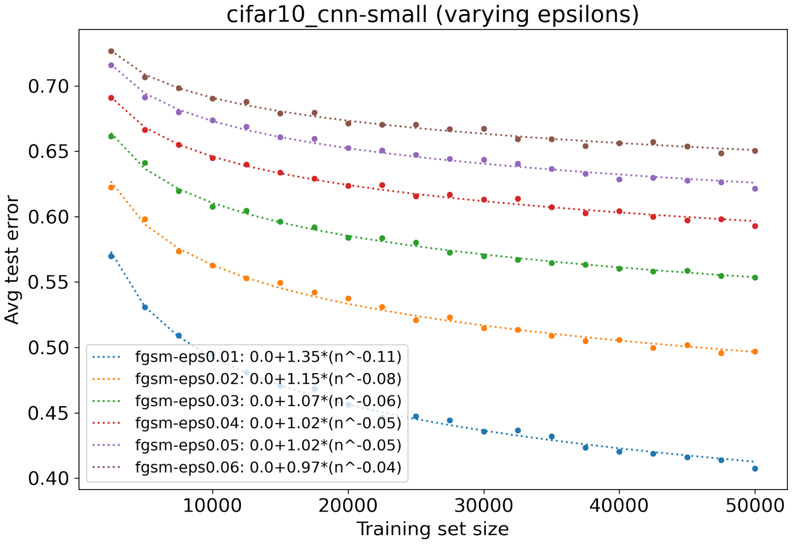


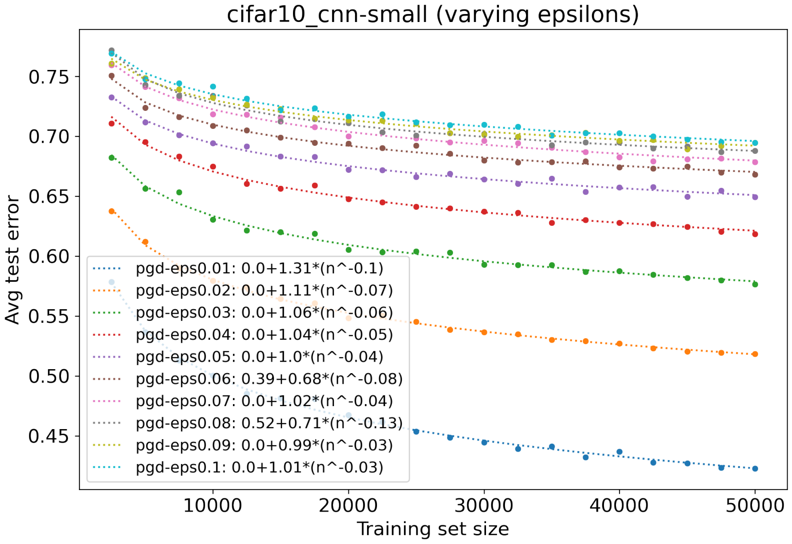
CIFAR-10 (Standard Accuracy)

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

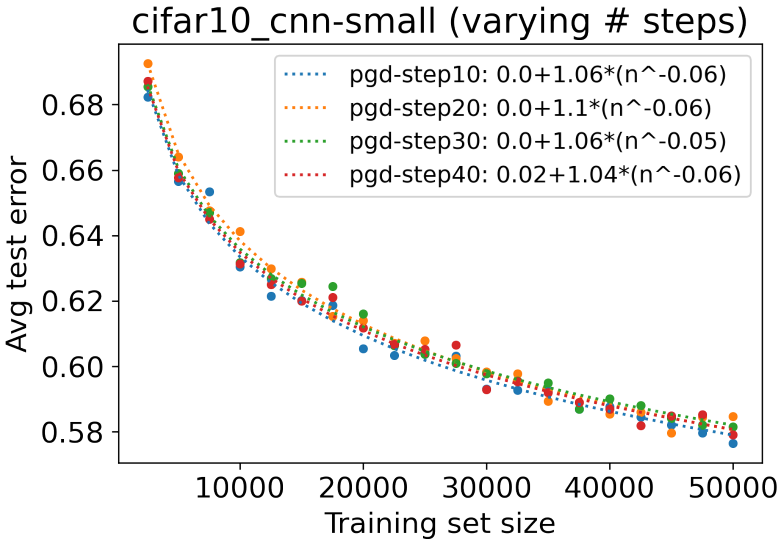
Table 3: Number of training examples required for adversarial training to reach the same standard accuracy as standard training on 100 examples using CIFAR-10.

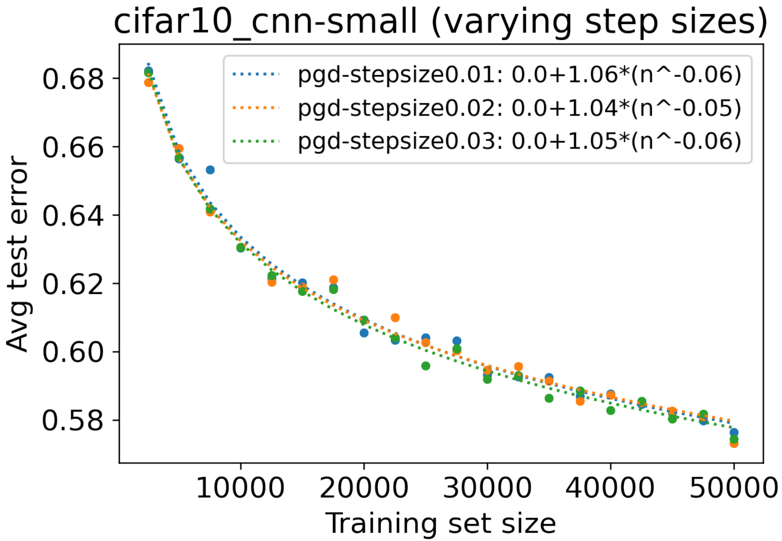
Increasing Epsilon





Increasing # Steps and Step Size





Conclusions

1. Both FGSM and PGD make the empirical sample complexity rate slower, requiring **up to ~10^5 times more data** to achieve the robust accuracy in standard training using MNIST and CIFAR-10 with MLP and various CNN architectures.
2. It is also the case for standard accuracy, but only **up to ~10 times** the sample size.
3. **Larger epsilon** leads to slower sample complexity rate, whereas increasing # steps or step size does not affect the rate.