Adversarially Robust Generalization Requires More Data

Ludwig Schmidt

Berkeley
UNIVERSITY OF CALIFORNIA

Shibani Santurkar

Dimitris Tsipras

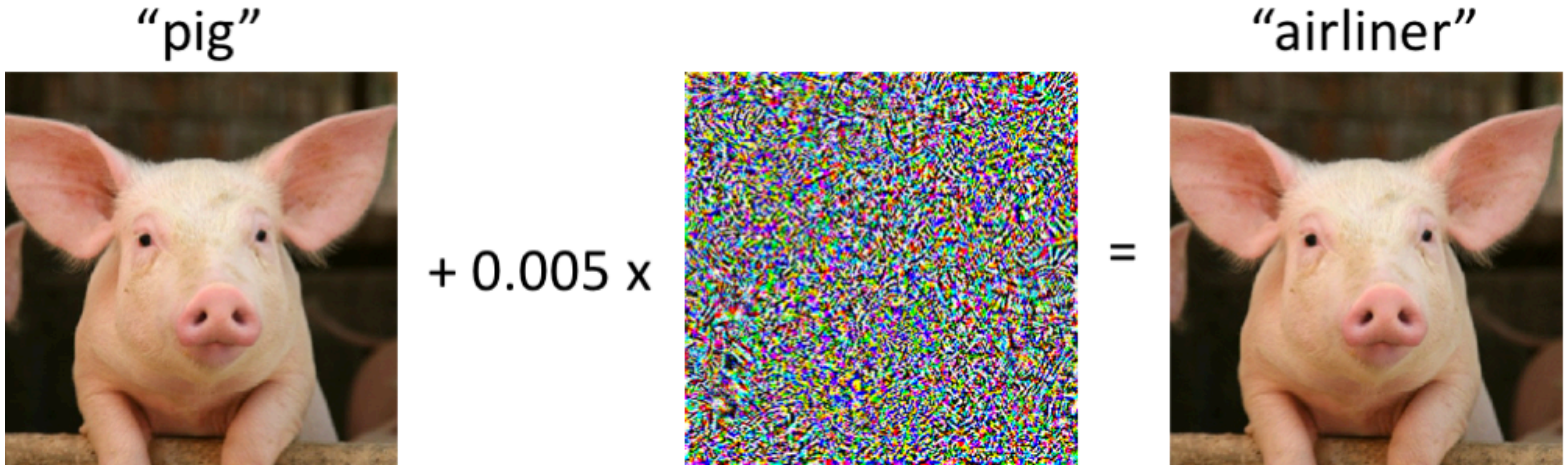
Kunal Talwar

Google

Aleksander Mądry

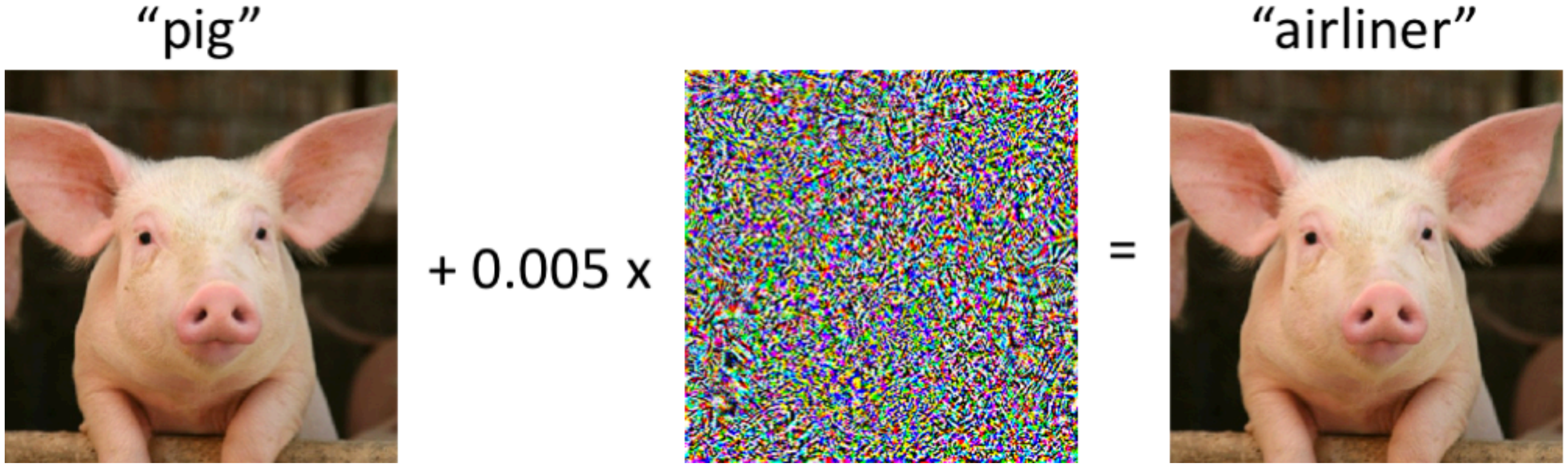
Poster #31

Adversarial Examples



[Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus, 2013]
[Biggio, Corona, Maiorca, Nelson, Srndic, Laskov, Giacinto, Roli, 2013]

Adversarial Examples



[Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus, 2013]
[Biggio, Corona, Maiorca, Nelson, Srndic, Laskov, Giacinto, Roli, 2013]

What makes adversarial examples a hard problem?

This paper: perspective on **sample complexity**

Standard vs Robust Generalization

"Standard" Generalization: $\underset{x,y \sim \mathcal{D}}{\mathbb{E}} \left[\text{loss}(f(x), y) \right]$

Standard vs Robust Generalization

"Standard" Generalization: $\mathbb{E}_{x,y\sim\mathcal{D}}\left[\,\text{loss}(f(x), y)\,\right]$

Adversarially robust generalization: $\mathbb{E}_{x,y\sim\mathcal{D}}\left[\max_{x'\in P(x)} \text{loss}(f(x'), y)\right]$

Perturbation set: small ℓ_∞-perturbations, rotations, translations, …

Standard vs Robust Generalization

"Standard" Generalization: $$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\,\text{loss}(f(x),y)\,\right]$$

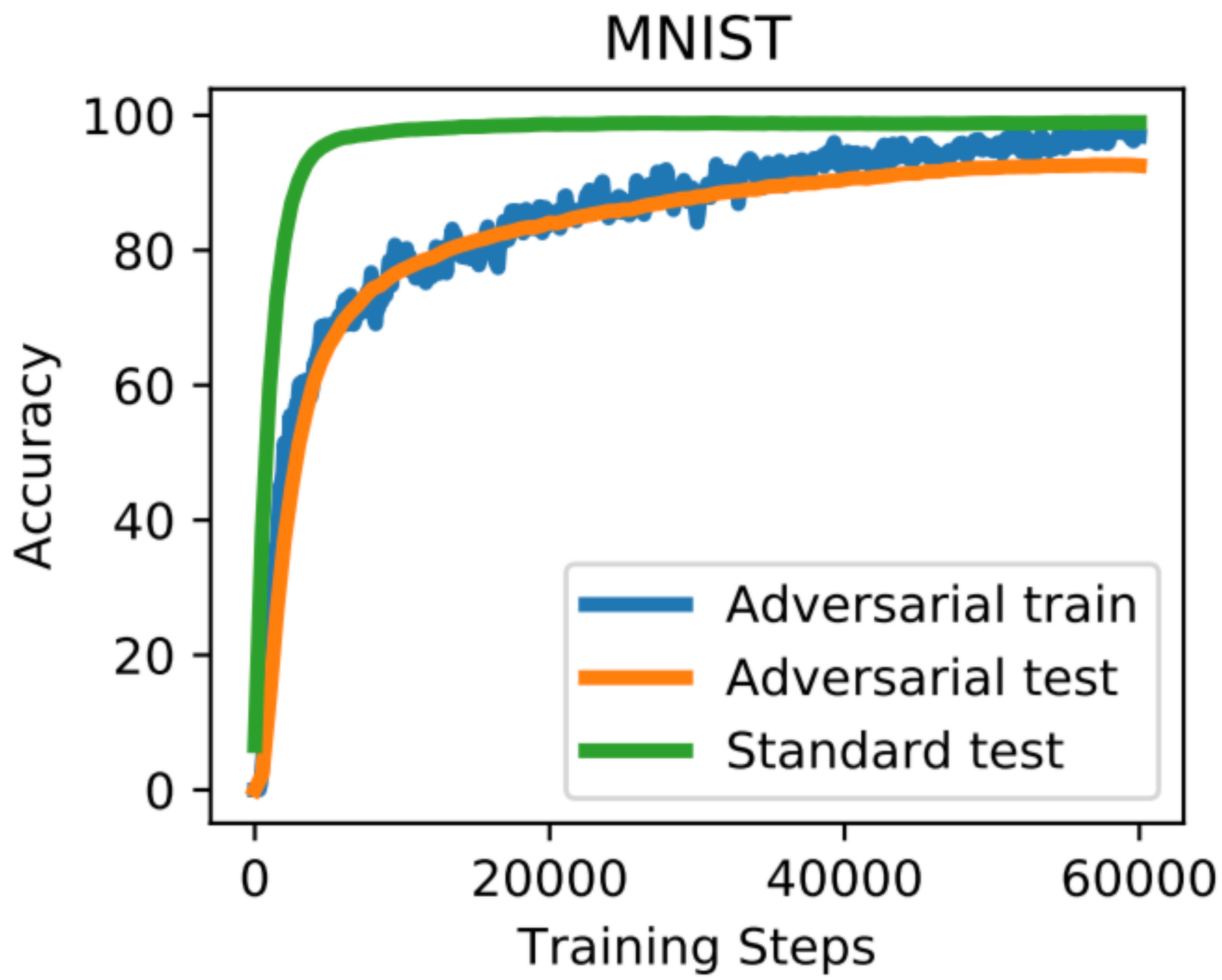
Adversarially robust generalization: $$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\max_{x'\in P(x)}\text{loss}(f(x'),y)\right]$$

Perturbation set: small ℓ_∞-perturbations, rotations, translations, …

How do these two notions of generalization compare?

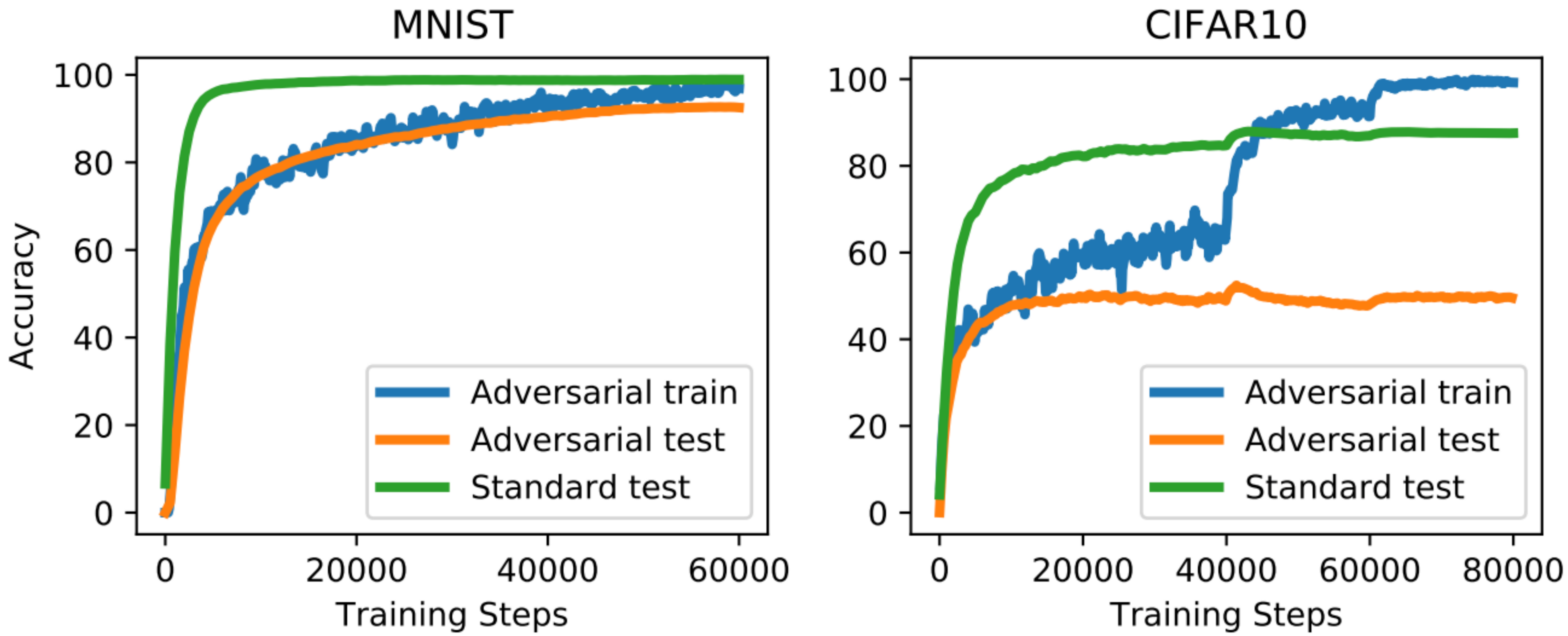
State Of The Art in ℓ_∞-Robustness

Robust optimization as in [Madry, Makelov, Schmidt, Tsipras, Vladu, 2017]:



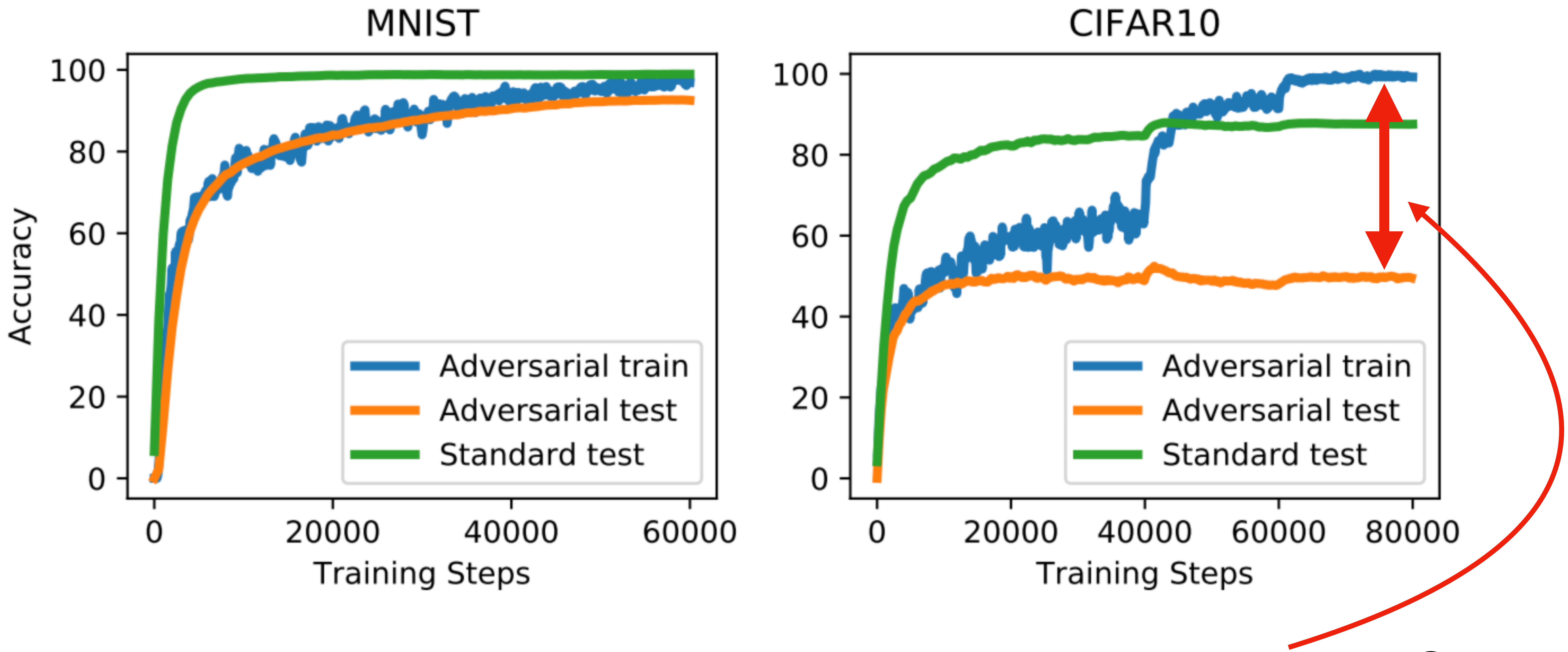
State Of The Art in ℓ_∞-Robustness

Robust optimization as in [Madry, Makelov, Schmidt, Tsipras, Vladu, 2017]:



State Of The Art in ℓ_∞-Robustness

Robust optimization as in [Madry, Makelov, Schmidt, Tsipras, Vladu, 2017]:



Optimization succeeds in both cases, but the model **overfits** on CIFAR-10.

Robust Generalization

Main question: Does robust generalization require more data?

Robust Generalization

Main question: Does robust generalization require more data?

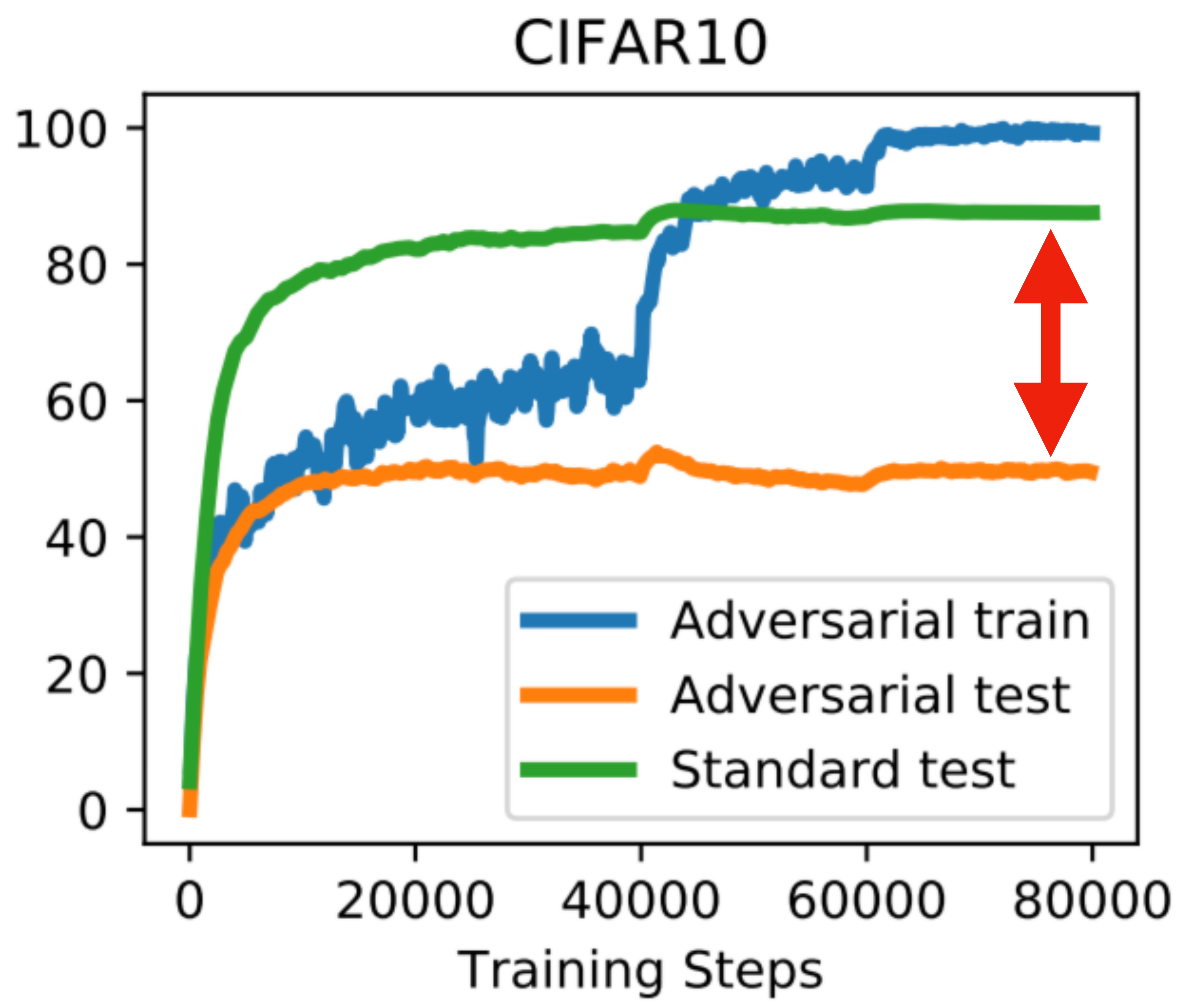
Theorem (informal): There is a natural distribution over points in R^d with the following property:

Learning an ℓ_∞-robust classifier for this distribution requires $\sqrt{d}$ times more samples than learning a non-robust classifier.

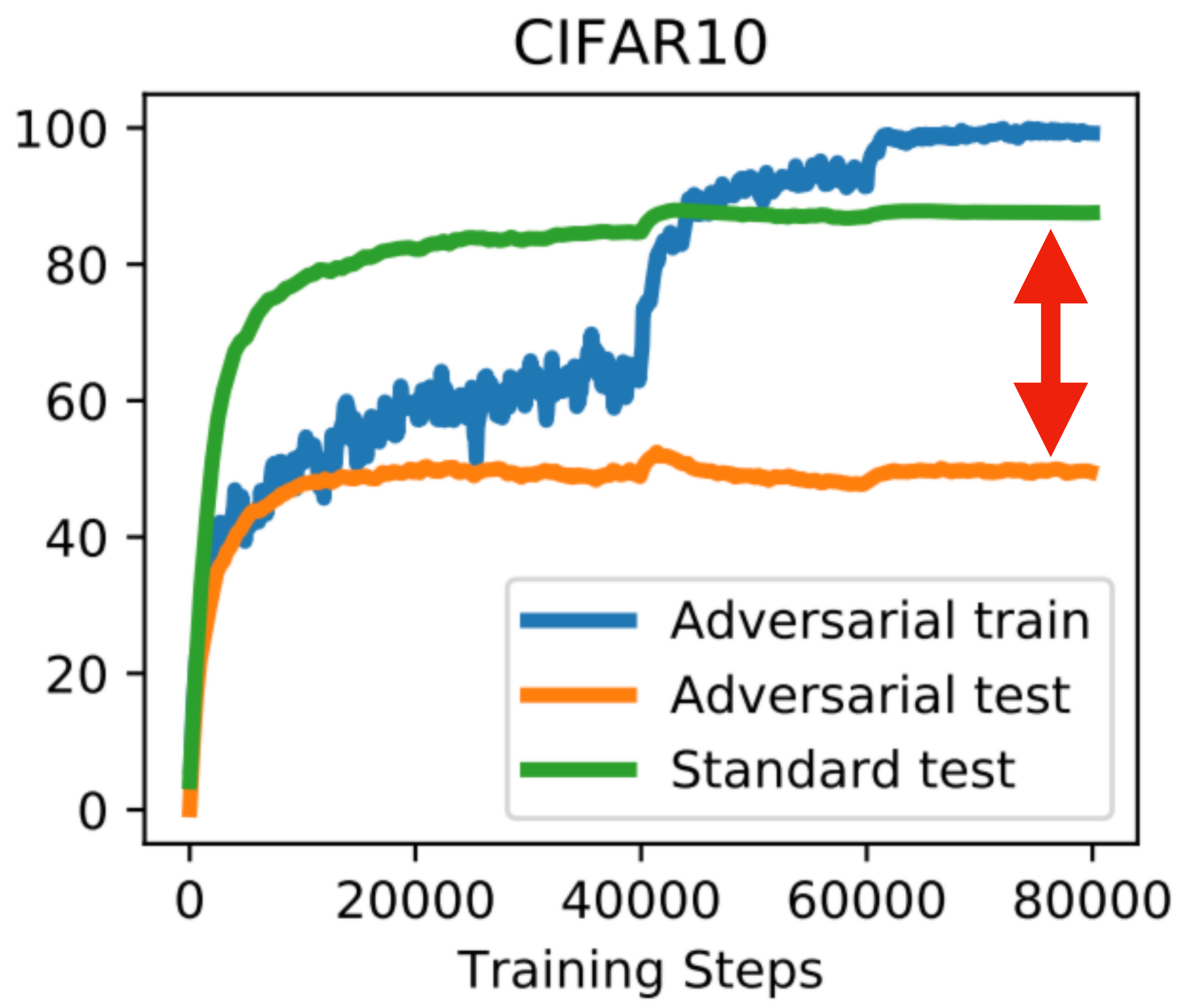
Conclusions

Further results

- An alternative data model for MNIST
- Experiments on MNIST, CIFAR-10, SVHN



Conclusions



Further results

- An alternative data model for MNIST
- Experiments on MNIST, CIFAR-10, SVHN

Main takeaways

- **Sample complexity** can be an obstacle for adv. robustness
- Need to **improve priors** encoded in models?
- Many phenomena not yet understood theoretically

Poster #31

gradient-science.org