Adversarial Robustness Guarantees for Random Deep Neural Networks

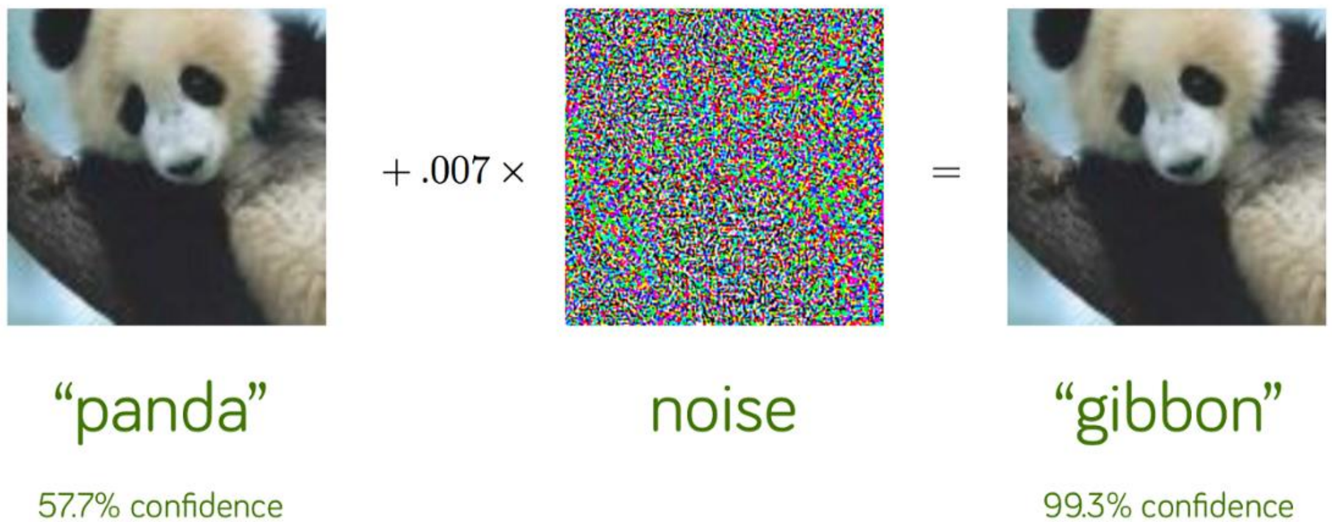
Giacomo De Palma

Bobak Kiani

Seth Lloyd

Adversarial examples

- Adversarial perturbation: extremely small perturbation that changes label of correctly classified input



- Challenge reliability of deep learning algorithms
- Still poor theoretical understanding

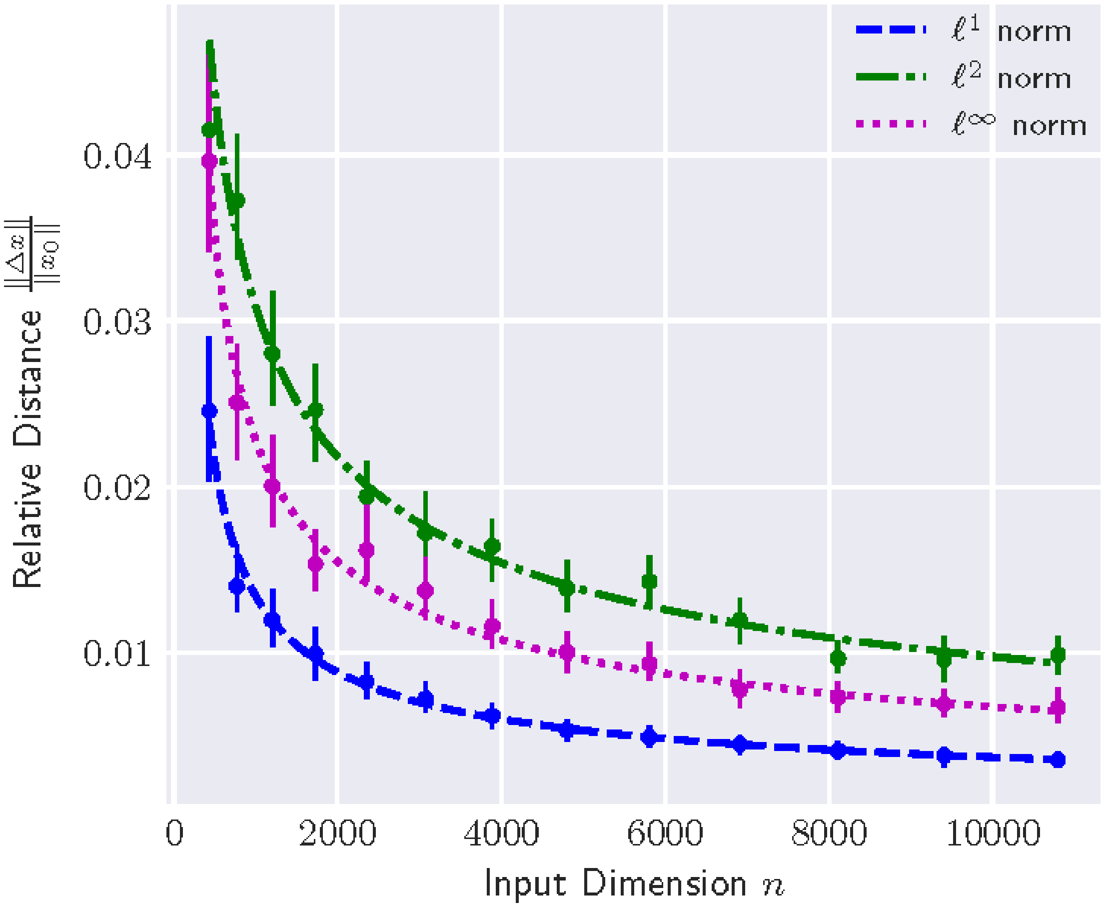
Adversarial Robustness Guarantees

- Independent random weights and biases
- Infinite width limit
- For any input x with entries with $O(1)$ magnitude and any $p \geq 1$, with high probability the ℓ^p distance to the classification boundary is at least

$$\boxed{d_p \geq \tilde{\Omega}\left(\frac{\|x\|_p}{\sqrt{n}}\right)} \qquad \|x\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

- Applies to any combination of fully connected or convolutional layers, skipped connections and pooling
- Applies to DNNs trained with Bayesian inference if target function generated by random DNN employed as prior

Experiments on random convolutional DNNs (7 hidden layers)



Trained convolutional DNNs

- MNIST: training does not change distance to boundary
- CIFAR10: training decreases distance to boundary due to visual structure (background, relevant part can be small)