Learning ReLU networks to high uniform accuracy is intractable

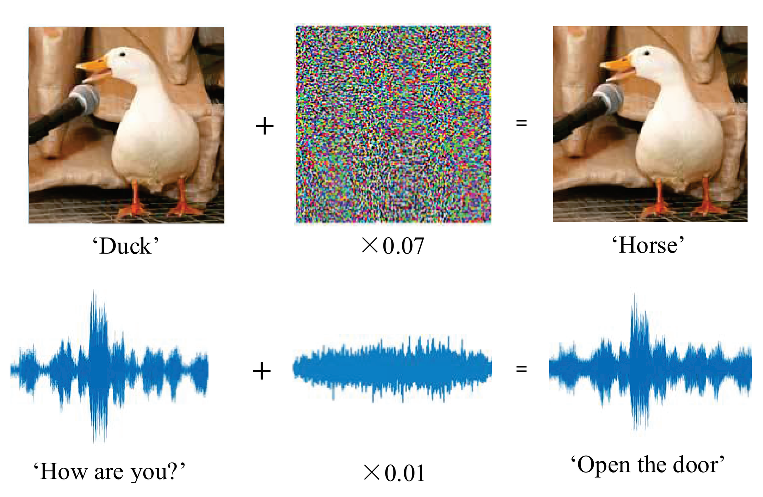
Julius Berner[1], **Philipp Grohs**[1,2], **Felix Voigtlaender**[3]

[1]University of Vienna, [2]RICAM, [3]Catholic University of Eichstätt-Ingolstadt

Instabilities in Deep Learning

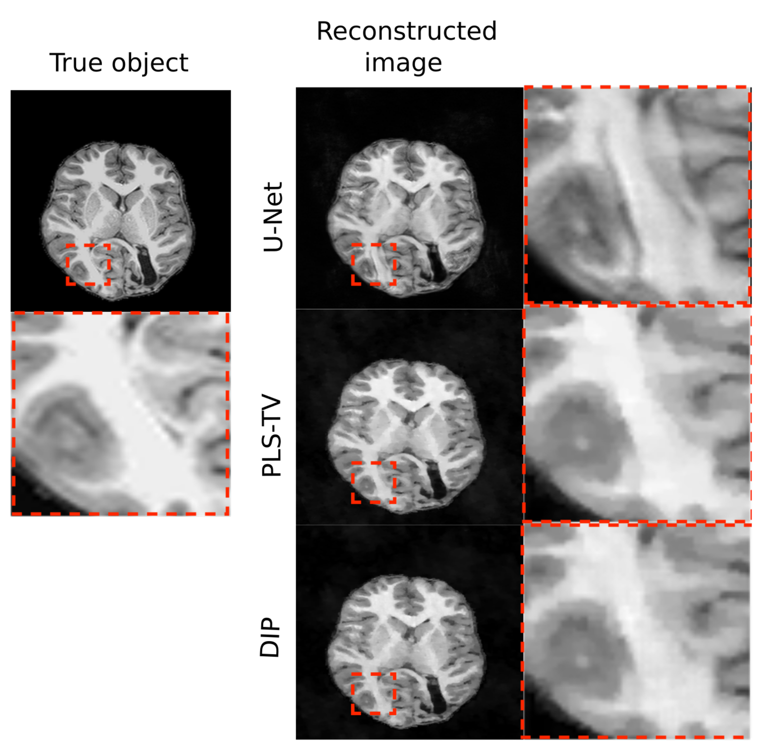
Undesired outputs of trained neural networks, even for inputs within the training distribution.

⚠ Adversarial examples



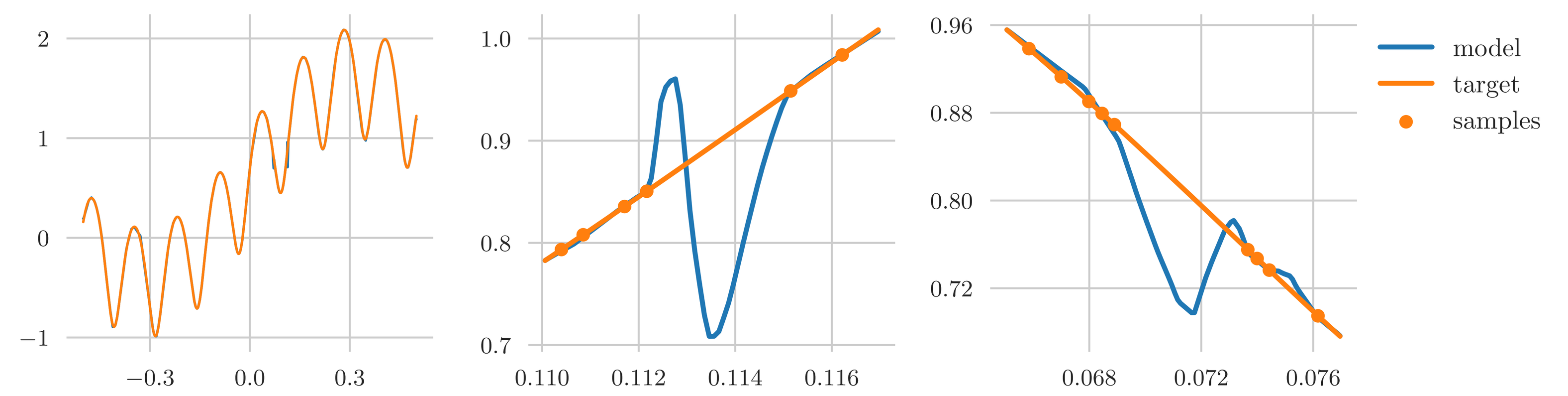
Y. Gong and C. Poellabauer. Protecting voice controlled systems using sound source identification based on acoustic cues. In ICCCN, 2018

⚠ Hallucinations



S. Bhadra et al. On hallucinations in tomographic image reconstruction. IEEE transactions on medical imaging, 2021

⚠ Inaccurate function approximations



Learning Theory

Generalization results only provide **guarantees in an average sense** (w.r.t. the L^2-norm).

👍 Approximation: Neural networks $\mathcal{N}$ can optimally approximate many function classes U (w.r.t. the L^∞-norm) in terms of the number of parameters required to guarantee

$$\sup_{u\in U}\inf_{f\in\mathcal{N}}\|f-u\|_{L^\infty}\le\varepsilon.$$

👍 Generalization: Bounds on the number of samples m required for the empirical risk minimizer $\hat f\in\arg\min_{f\in\mathcal{N}}\sum_{i=1}^m(f(x_i)-y_i)^2$ to approximate the optimal neural network f^* (w.r.t. the L^2-norm), i.e.,

$$\|\hat f-f^*\|_{L^2}\le\varepsilon,$$

often scale only polynomially in the underlying dimension d.

Our Lower Bound

Learning ReLU networks to high **uniform accuracy** (w.r.t. the L^∞-norm) **requires an intractable number of samples**.

We consider all algorithms $\mathcal{A}$ that operate on samples

$$(x_i,u(x_i))_{i=1}^m.$$

This includes:

- ✔ all variants of (S)GD,
- ✔ adaptive algorithms (e.g., active learning),
- ✔ randomized algorithms (e.g., MC algorithms),
- ✔ intractable algorithms (e.g., empirical risk minimization).

Assume that $\mathcal{N}\subset U$ consists of ReLU networks with input dimension d, $L\ge 3$ layers, width $3d$, and parameters bounded by c. Any algorithm $\mathcal{A}$ satisfying

$$\sup_{u\in U}\mathbb{E}\left[\|\mathcal{A}(u)-u\|_{L^\infty}\right]\le\varepsilon$$

requires

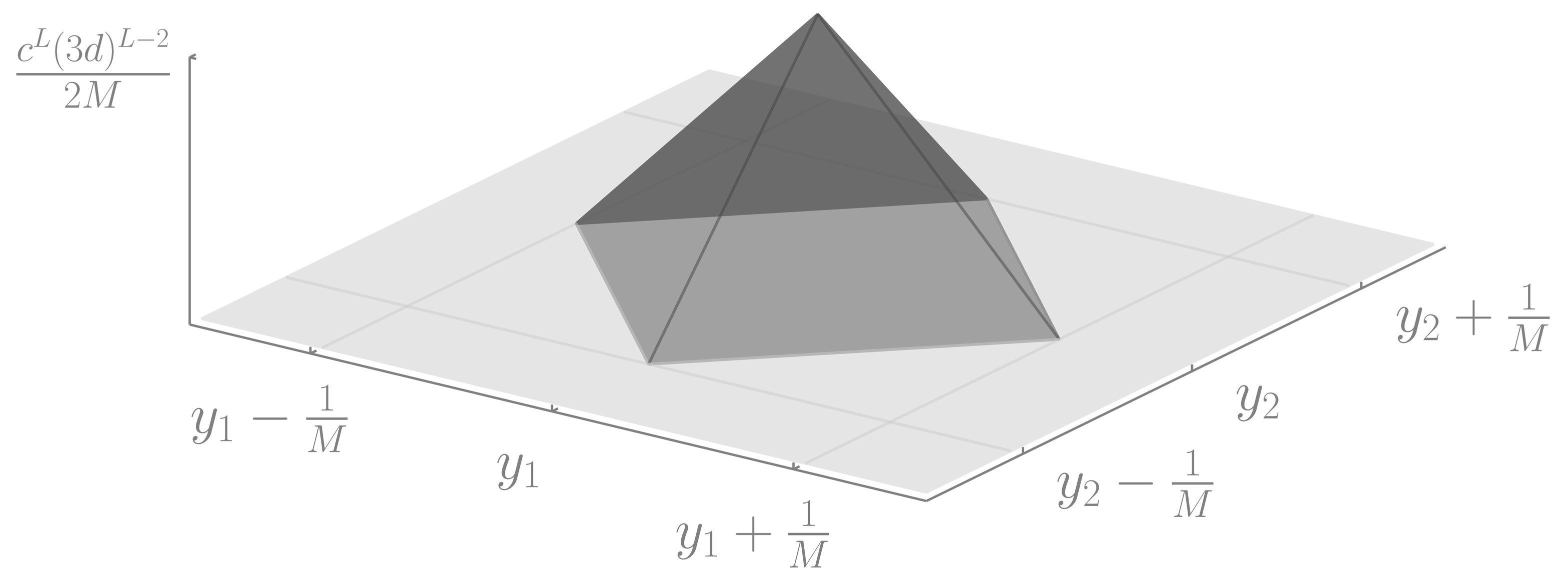
$$m\ge c^{dL}(3d)^{d(L-2)}\left(\frac{1}{2^9\varepsilon}\right)^d$$

samples on average.

👎 Number of samples m required to achieve high uniform accuracy ε scales exponentially with the underlying dimension d and the depth L of the ReLU networks $\mathcal{N}$.

👎 Different from other hypothesis classes (e.g., polynomials or certain RKHS), m can significantly exceed the number of parameters defining the class $\mathcal{N}$.

Proof idea: $\mathcal{N}\subset U$ contains localized bump functions f with $f(x_i)=0$ for all $i\in\{1,\dots,m\}$, such that $\mathcal{A}(\pm f)=\mathcal{A}(0)$.



Our Upper Bound

Our bounds are **asymptotically sharp**.

There exists an algorithm $\mathcal{A}$ (based on piecewise constant interpolation) that satisfies $\sup_{u\in\mathcal{N}}\mathbb{E}\left[\|\mathcal{A}(u)-u\|_{L^\infty}\right]\le\varepsilon$ using

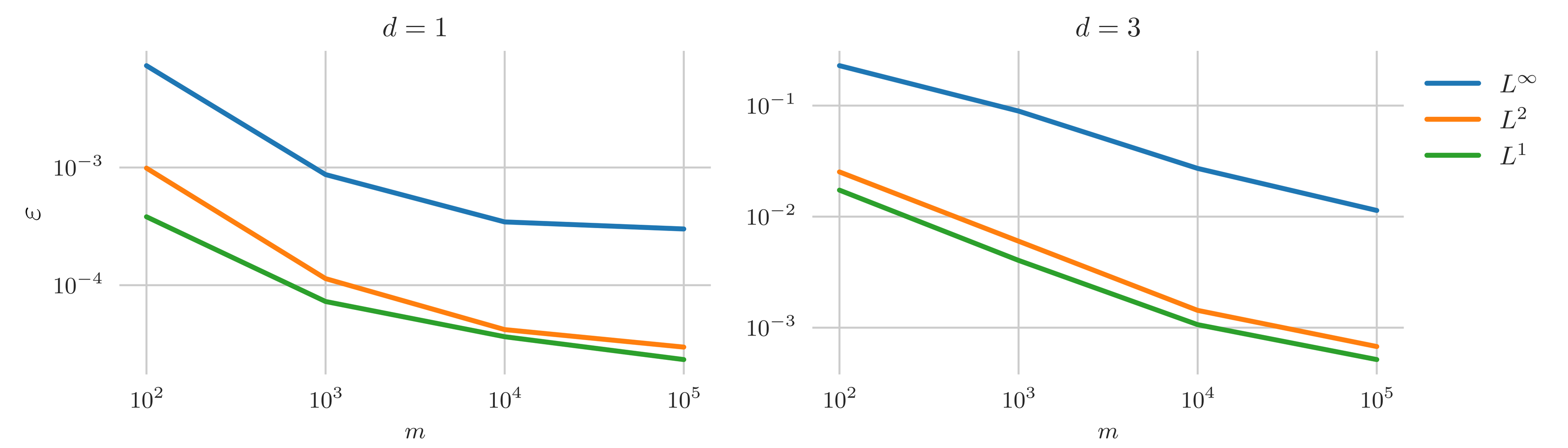
$$m\le c^{dL}(3d)^{d(L-2)}\left(\frac{3d^2}{\varepsilon}\right)^d$$

samples.

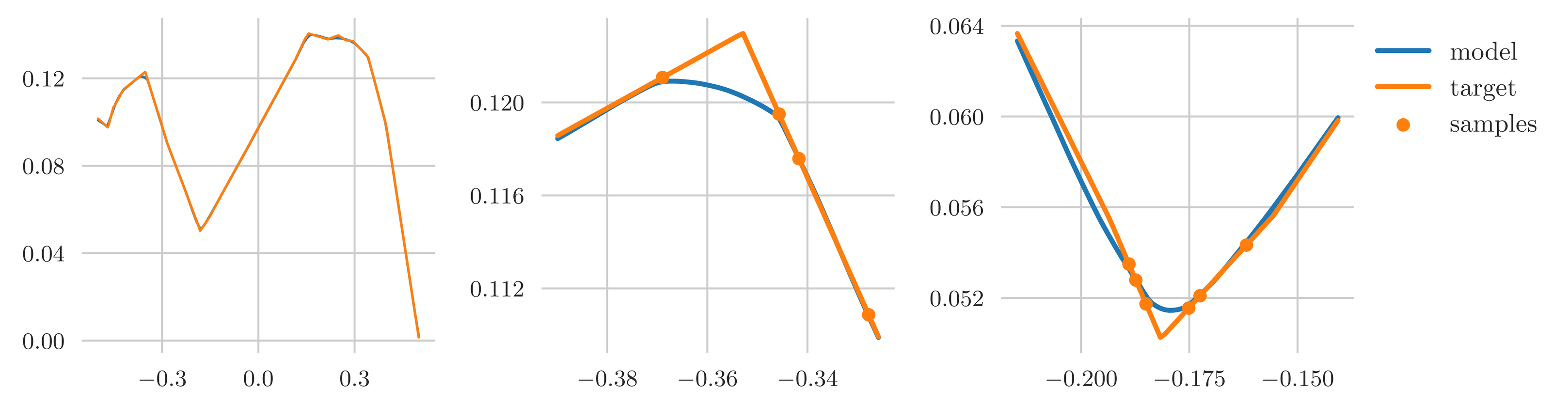
Numerical Experiments

Theoretical results are **validated in student-teacher settings**.

✔ Gap between uniform and average errors:



✔ Similar bumps prevent high uniform accuracies:



References and Further Results

```
arxiv.org/abs/2205.13531
github.com/juliusberner/theory2practice
```

- ✔ Fully explicit lower bounds for all L^p-norms and different parameter regularizations.
- ✔ Connections to statistical query algorithms, statistical learning theory, and neural network identification.