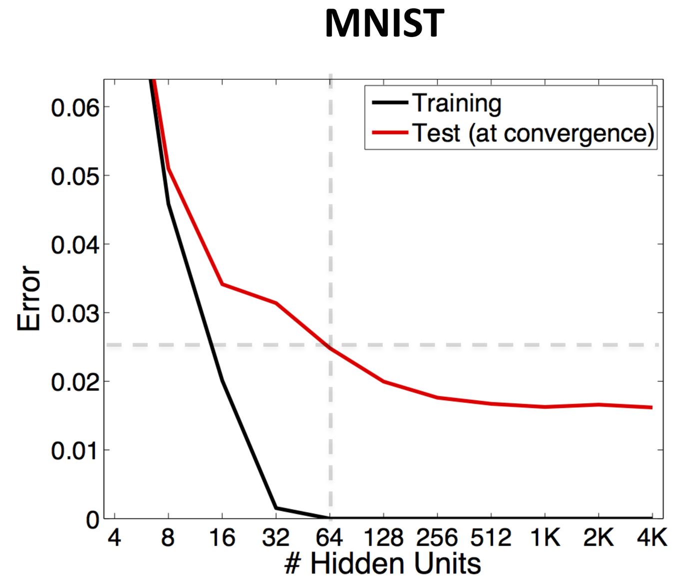
In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning

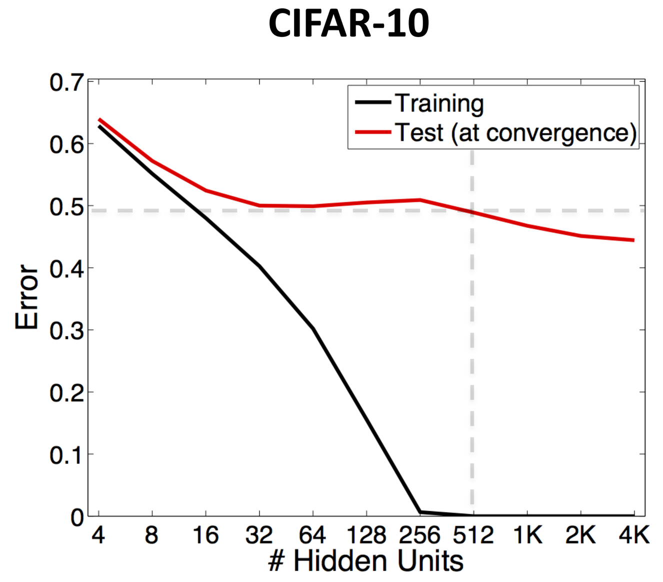
Behnam Neyshabur, Ryota Tomioka, Nathan Srebro.
TTI-Chicago

Network Size as Capacity Control?

A simple experiment: Learning a feed-forward network with a single hidden layer without any regularization.

We expect: Network Size↑⟹ Approximation error ↓, Estimation Error↑





Without any regularization, even with zero training error (and zero approximation error), increasing the number of hidden units **reduces** estimation error.

How can we explain this phenomenon?

Network Size as Inductive Bias?

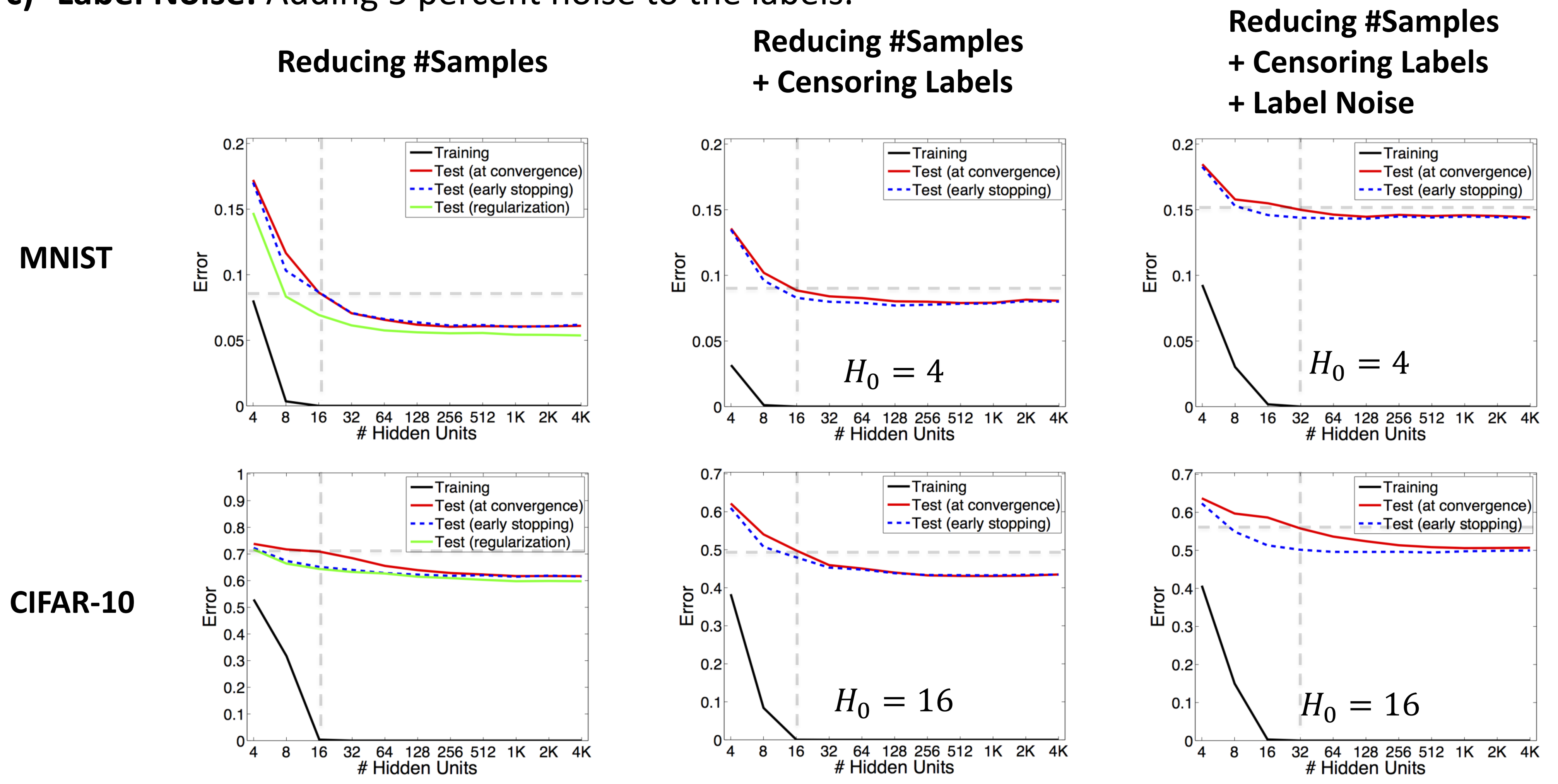
The fact that we can present data with a small network is NOT enough to ensure that we can learn it efficiently.

- **Computationally intractable even for small networks:** Even if target can be exactly represented by a network with a single hidden layer and $\log(d)$ units, no poly-time for learning even using much larger networks (or any other representation).☹
- **Being representable by NN is not enough as an inductive bias:** In fact, any $O(T)$ time computable function can be represented by $O(T^2)$ size network (Sipser, 2006).☺

Why do we succeed in learning using neural networks?
What property (inductive bias) makes them possible to learn?

Implicit Regularization

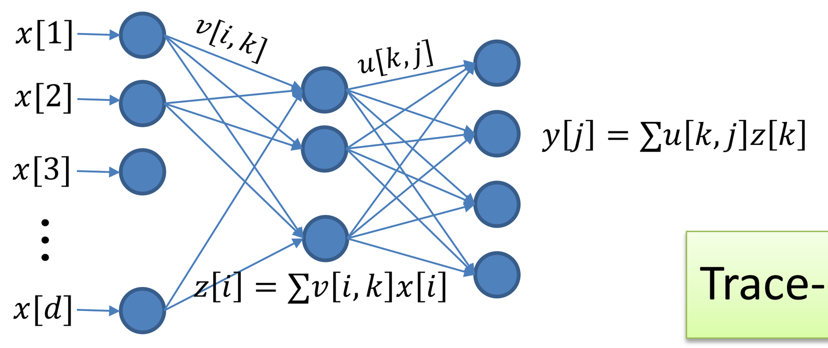
a) **Reducing #samples:** Reducing the size of training set to **2000**.
b) **Censoring Labels:** Switching the labels to the predictions of a network with a small number H_0 of hidden units that is trained on the entire dataset (training+validation+test).
c) **Label Noise:** Adding 5 percent noise to the labels.



What is happening here? A possible explanation:
Implicit regularization introduced by the optimization.
Converging to a global optima with "low complexity", perhaps low norm.

Matrix Factorization Analogy

Insights from a simpler model: linear activations



$$W = UV$$

$$y = U \cdot \underbrace{Vx}_{z} = Wx$$

r hidden units ⟺ $rank(W) \le r$

Trace-norm as an inductive bias: $\|W\|_{tr} = \min_{W=UV} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2)$

Matrix Factorization	Low r: intractable	Trace-norm	Higher rank ⇒ lower trace-norm⇒better generalization
Feed-forward Networks	Low r: intractable	Some norm?	More hidden units⇒ lower norm⇒better generalization?

Minimum norm: Infinite-sized networks?

Infinite Size, Bounded Norm Networks

Theorem 1. For a feed-forward network with a single hidden layer, weight decay (i.e. regularizing $\sum_{e\in E} w(e)^2$) is equivalent to bounding the L2 norm of the incoming weights to each hidden unit and regularizing the L1 norm of the incoming weights to the output unit.

Corollary. As long as $r > \#samples$, **weight decay regularized network is equivalent to convex NN (Bengio et al., 2005)**

Could we bound the capacity of a bounded-norm network with infinite number of hidden units?

Group-norm regularization. For any directed graph G, consider the following norm:

$$L_{p,q}(G) = \left(\sum_{v\in V}\left(\sum_{(u\to v)\in E} |w(u\to v)|^p\right)^{\frac{q}{p}}\right)^{1/q}$$

Theorem 2. For any $1 \le p \le 2$, if $\frac{1}{p} + \frac{1}{q} \ge 1$, can bound the sample complexity required for learning, even if unbounded (infinite) number of units, as long as $L_{p,q}$ bounded, as:

$$\text{sample complexity} \propto \left(\frac{2L_{p,q}}{d}\right)^{2d}$$

and this is tight up to multiplicative factors multiplying d (i.e. up to replacing d with Cd for some constant C).

Examples:
- $p = q = 2$ weight decay
- $p = q = 1$ overall sum of absolute weights
- $p = 1, q = \infty$ per unit ℓ_1 norm
- If $\frac{1}{p} + \frac{1}{q} < 1$, class of NN of depth ≥ 3 with unbounded #units and bounded $L_{p,q}$ has infinite capacity.

Norm-Based Capacity Control in Neural Networks.
Behnam Neyshabur, Ryota Tomioka, Nati Srebro.
The 28th Conference on Learning Theory (COLT), 2015 (to appear).