A PAC-Bayesian Approach to Generalization in Deep Learning

Behnam Neyshabur

Institute for Advanced Study

Joint work with Srinadh Bhojanapalli, David McAllester, Nati Srebro

Observations about Neural Nets

- Deep networks are *over-parametrized:*

  #parameters >> #samples

- Many global optima
  - Some of them do not generalize well!

- Choice of optimization $\Rightarrow$ different global minimum $\Rightarrow$ different generalization

Requirements for a complexity measure that explains generalization

w: the parameter vector.

$R(w)$: complexity measure, ex. $R(w) = \|w\|_2$

1. $\{w|R(w) \text{ is small}\}$ has small capacity, i.e. small R(w) is sufficient for generalization
2. Natural problems can be predicted by $\{w|R(w) \text{ is small}\}$
3. The optimization algorithm biases us towards solutions in $\{w|R(w) \text{ is small}\}$

Outline

- From PAC-Bayes to Margin
- From PAC-Bayes to Sharpness
- Empirical Investigation of three phenomena:
  - Fitting random labels (Zhang et al., 2016).
  - Different global minima (Keskar et al., 2016).
  - Large networks generalize better (Neyshabur et al., 2015).

Preliminaries

- Feedforward nets: $f_{\mathbf{w}}(\mathbf{x}) = W_d\,\phi(W_{d-1}\,\phi(....\phi(W_1\mathbf{x})))$
  - d layer
  - h hidden unit in each layer
  - ReLU activations $\phi(x) = \max\{0, x\}$
  - B bound on ℓ_2-norm of x

- Margin Loss:

$$L_\gamma(f_w) = P_{(x,y)}[\text{score of } y - score\ of\ other\ labels \leq \gamma]$$

  - Misclassification error: $L_0(f_w)$

Capacity Control

- Network Size
  - The capacity is too high.
  - Can’t explain any of the phenomena.

- Scale Sensitive Capacity Control:
  - Scale of the predictor, i.e. weights
  - Scale of the predictions (Margin or Sharpness)

Margin

$\gamma =$ *score of the correct label – maximum score of other labels*

Margin-based measures:

- ℓ_2-norm with capacity $\propto \frac{\prod_{i=1}^{d} \|W_i\|_F^2}{\gamma^2}$ (Neyshabur et al. 2015)
- ℓ_1-path norm with capacity $\propto \frac{\phi_{path,1}^2}{\gamma^2}$ (Bartlett and Mandelson 2002)
- ℓ_2-path norm with capacity $\propto h^d \frac{\|W_i\|_{path,2}^2}{\gamma^2}$ (Bartlett and Mandelson 2002)
- spectral norm with capacity $\propto \frac{\prod_{i=1}^{d} \|W_i\|_2^2}{\gamma^2} \left(\sum_{i=1}^{d} \frac{\|W_i\|_{1,2}^{\frac{2}{3}}}{\|W_i\|_2^{\frac{2}{3}}} \right)^3$ **(Bartlett et al. 2017)**

$\|.\|_F$: Frobenius norm $\|.\|_2$: Spectral norm $|.|_p$: ℓ_p norm of a vector $\|.\|_{path,p}$: ℓ_p-path norm

PAC-Bayes

Theorem (McAllester 98): For any P and any $\delta \in (0,1)$
w.p $1 - \delta$ over the choice of the training set S, for any Q:

$$\mathbb{E}_{\mathbf{w}\sim Q}[L_0(f_{\mathbf{w}})] \leq \mathbb{E}_{\mathbf{w}\sim Q}[\widehat{L}_0(f_{\mathbf{w}})] + \sqrt{\frac{KL\left(Q\|P\right) + \ln\frac{m}{\delta}}{2(m-1)}}$$

What if we want to get generalization for a given weight w?

- Consider the distribution over $w + u$ where u is random perturbation.

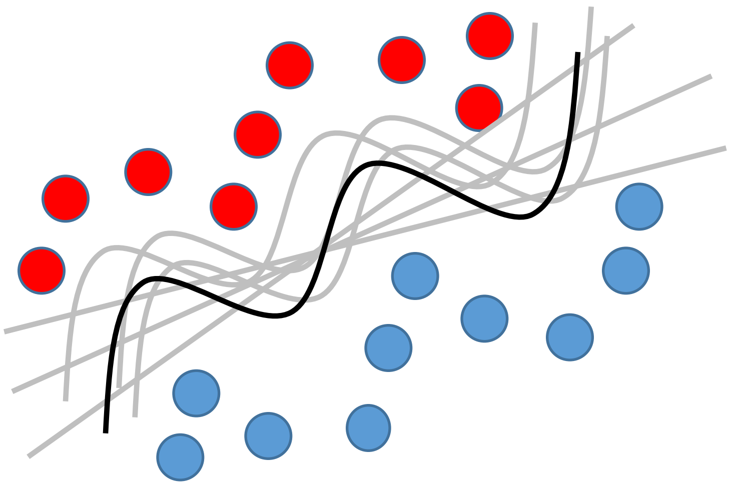
PAC-Bayes(2)

Theorem: For any P and any $\delta \in (0,1)$ w.p $1-\delta$ over the choice of the training set S, for any w and Q over u:

$$\mathbb{E}_{\mathbf{u}\sim Q}[L_0(f_{\mathbf{w}+\mathbf{u}})] \leq \mathbb{E}_{\mathbf{u}\sim Q}[\widehat{L}_0(f_{\mathbf{w}+\mathbf{u}})] + \sqrt{\frac{KL\left(\mathbf{w}+\mathbf{u}\|P\right)+\ln\frac{m}{\delta}}{2(m-1)}}$$

From margin to PAC-Bayes

Large margin: small perturbation in parameters will not change the loss.



From PAC-Bayes to margin

Lemma 1: For any P and any $\gamma > 0, \delta \in (0,1)$ w.p $1-\delta$ over the choice of the training set S, for any Q over u such that

$$\mathbb{P}_{\mathbf{u}\sim Q}\left[\max_{\mathbf{x}\in\mathcal{X}} |f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})|_\infty < \frac{\gamma}{4}\right] \geq \frac{1}{2}$$

we have:

$$L_0(f_{\mathbf{w}}) \leq \widehat{L}_\gamma(f_{\mathbf{w}}) + \sqrt{\frac{KL\left(\mathbf{w}+\mathbf{u}\|P\right) + \ln\frac{3m}{\delta}}{m-1}}$$

Proof idea: similar analysis for linear predictors (Langford & Shawe-Taylor (2003) and McAllester (2003)).

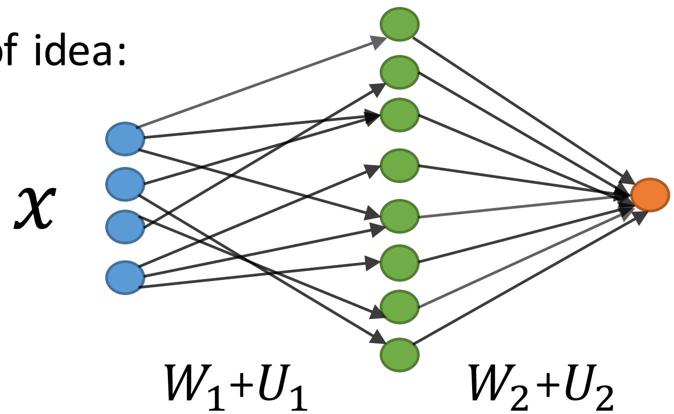
Perturbation Bound

How much the network output changes if we perturb the parameters?

Lemma 2: For any perturbation u such that $\|U_i\|_2 \leq \frac{1}{d}\|W_i\|_2$

$$|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})|_2 \leq eB\left(\prod_{i=1}^{d}\|W_i\|_2\right)\sum_{i=1}^{d}\frac{\|U_i\|_2}{\|W_i\|_2}.$$

Proof idea:



$$f_{w+u}(x) \approx f_w(x) + \langle \nabla f_w(x), u \rangle$$

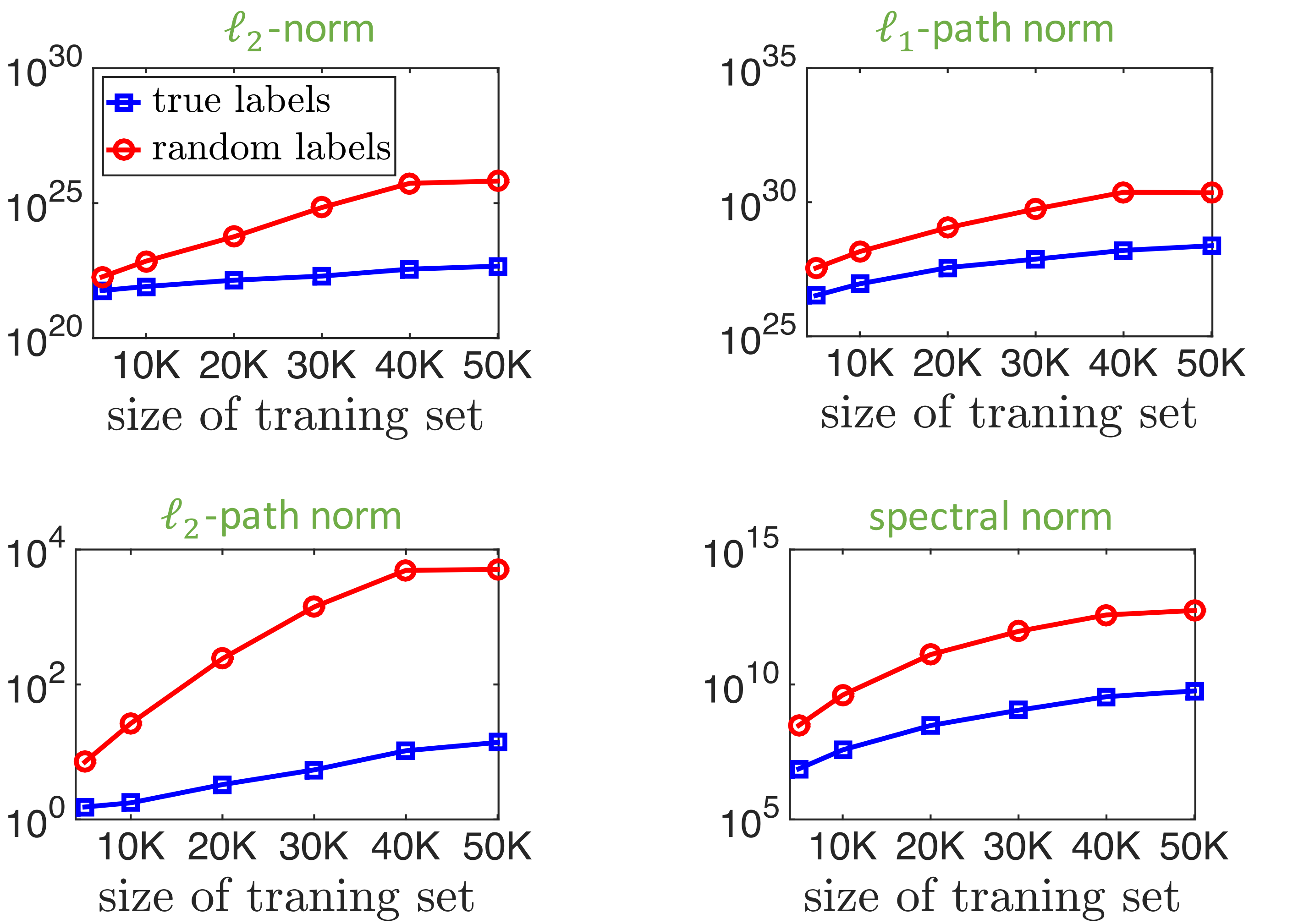
Generalization Bound for Neural Nets

Theorem: For any $\gamma > 0, \delta \in (0,1)$ w.p $1 - \delta$ over the choice of the training set

$$L_0(f_{\mathbf{w}}) \leq \widehat{L}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{d^2 h \ln(dh) B^2 \Pi_{i=1}^d \|W_i\|_2^2 \sum_{i=1}^d \frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \ln \frac{dm}{\delta}}{\gamma^2 m}}\right)$$

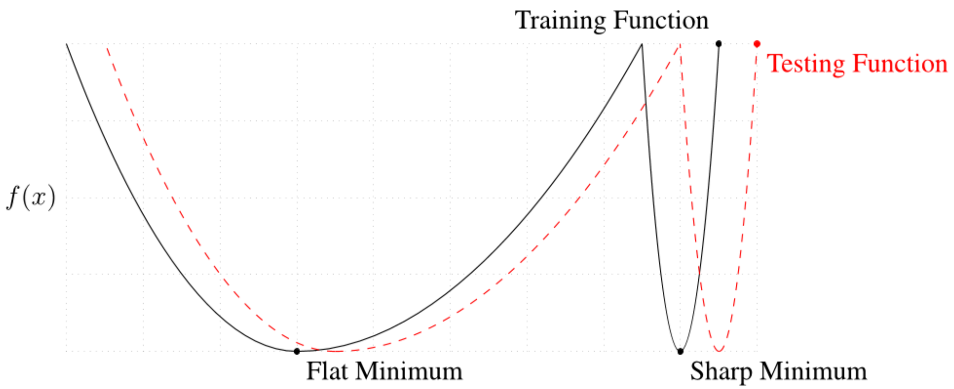
Proof idea: Choose prior and posterior both to be independent Gaussian distributions.

Experiments on True and Random Labels

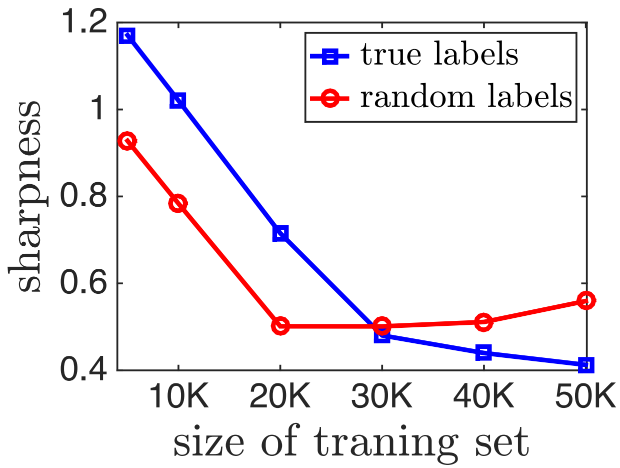


Sharpness

$$sharpness(\alpha) = \max_{\|\nu\| \leq \alpha} L(w+\nu) - L(w)$$ [Keskar et al.17]



Similar to margin, controlling sharpness alone is meaningless.

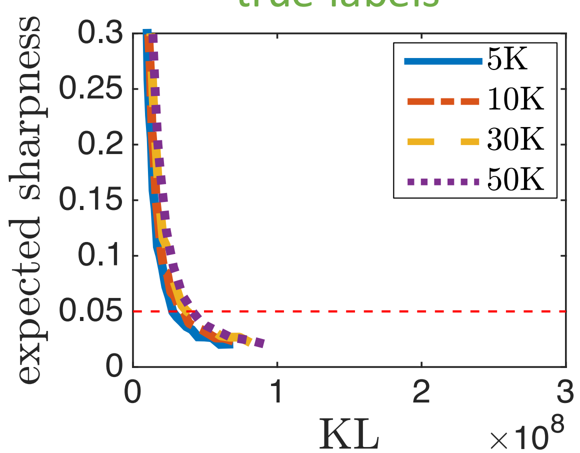


From PAC-Bayesian to Sharpness

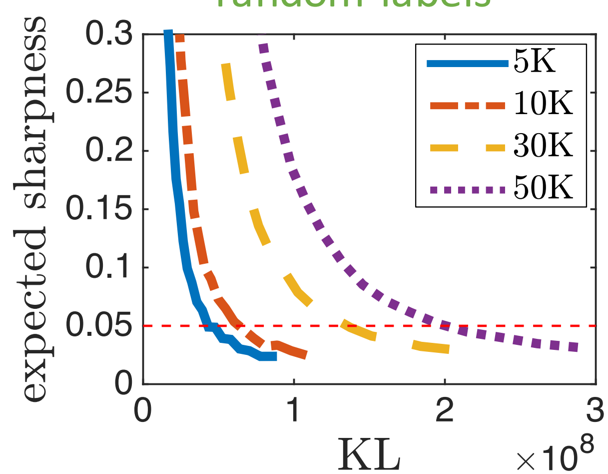
- Sharpness can be understood as one of the two terms in the PAC-Bayes bound (Dziugaite and Roy 2017).

$$\mathbb{E}_\nu[L(w+\nu)] \le \hat{L}(w) + \underbrace{\mathbb{E}_\nu\left[\hat{L}(w+\nu)\right] - \hat{L}(w)}_{\text{expected sharpness}} + \sqrt{\underbrace{\frac{1}{m}\left(KL(w+\nu||P) + \ln\frac{2m}{\delta}\right)}_{\frac{\|w\|_2^2}{2\sigma^2} \text{ if } \begin{cases} P = N(0,\sigma^2) \\ \nu \sim N(0,\sigma^2)\end{cases}}}$$

true labels

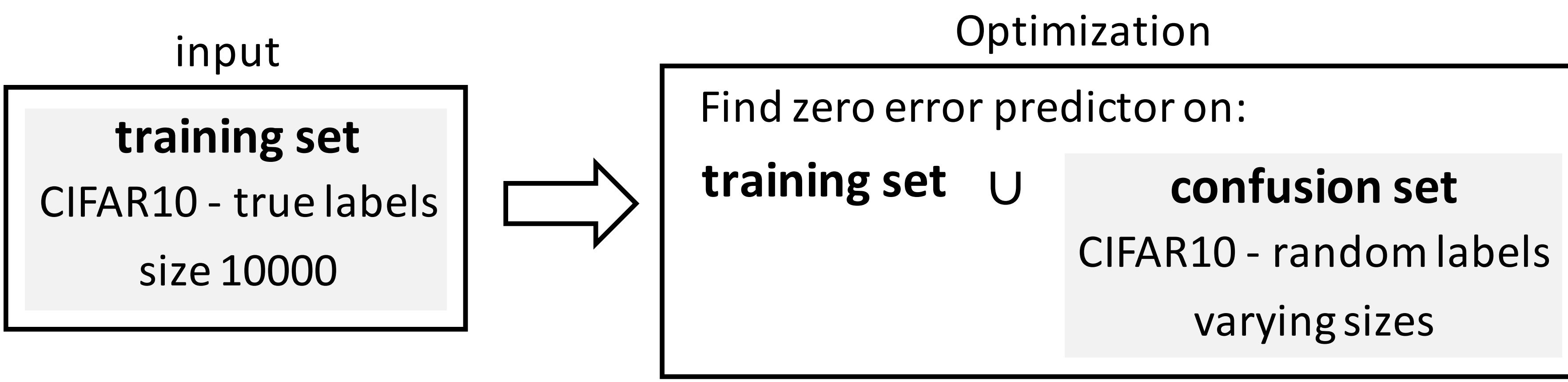


random labels

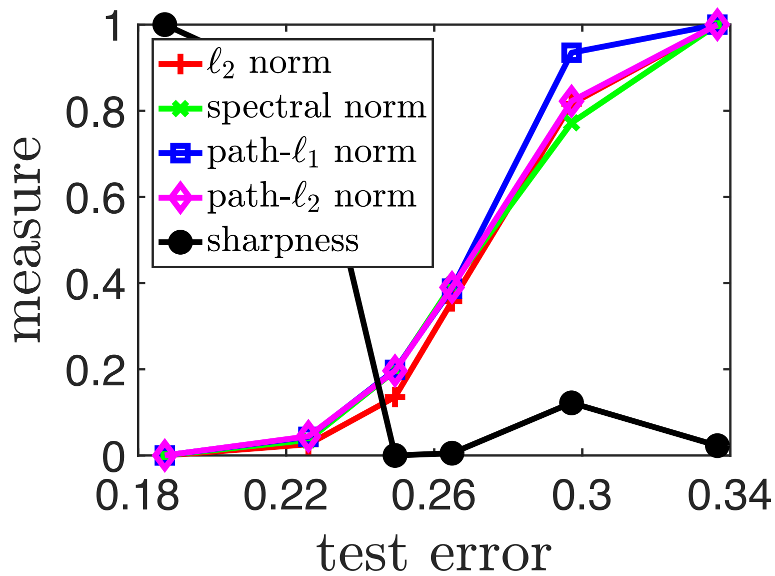


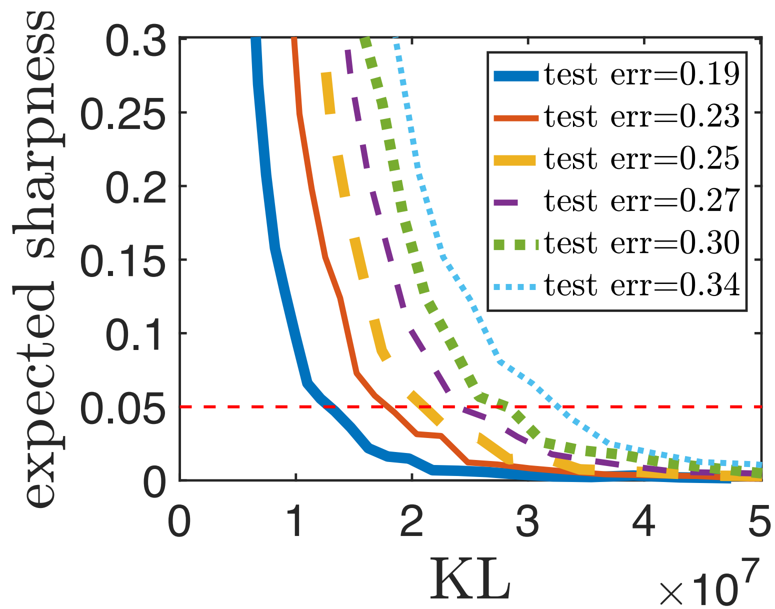
Generating Different Global Minima

- We construct different global minima of the training loss for the same data, intentionally with different generalization properties. How?

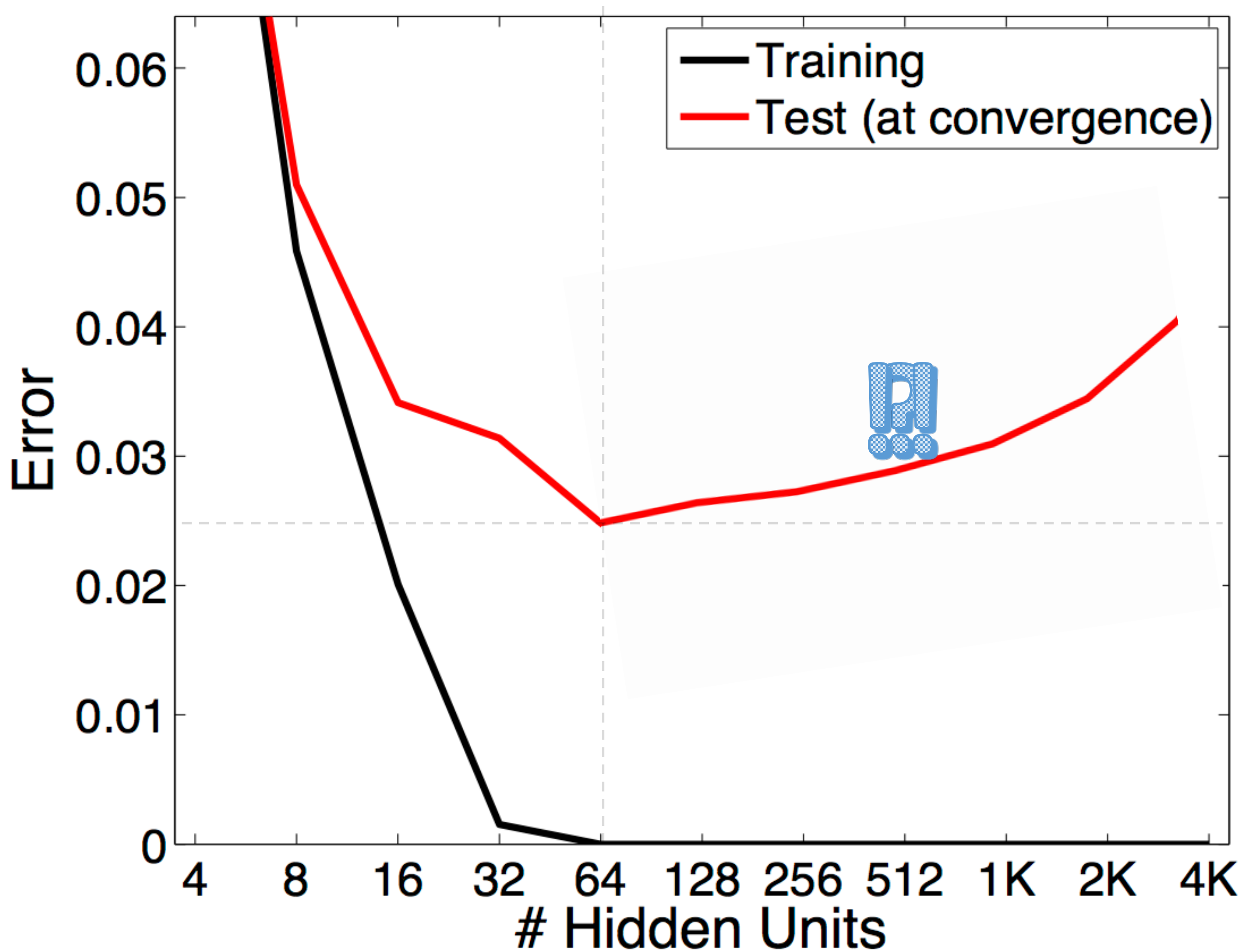


Different global minima



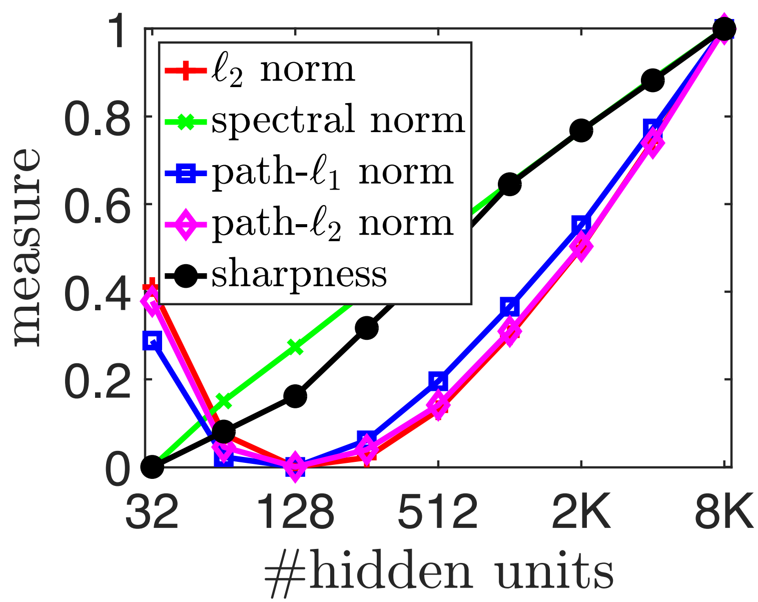


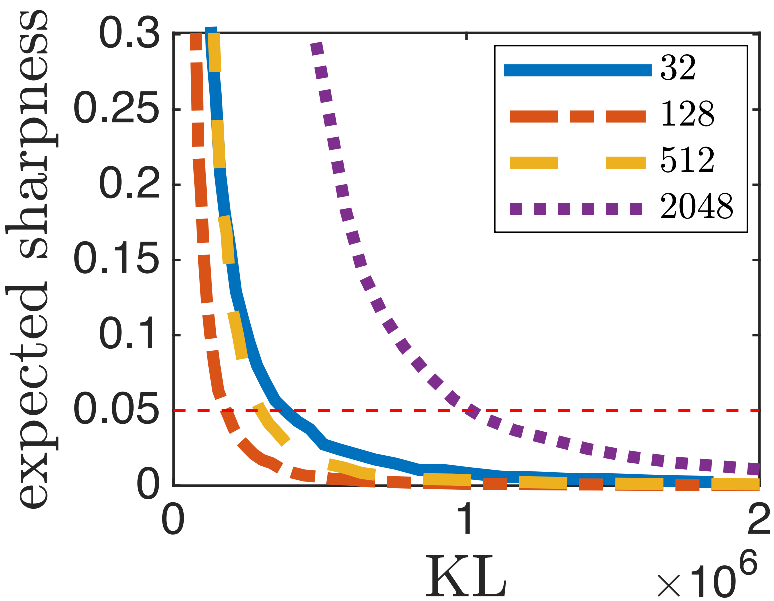
Increasing the Network Size (Number of Hidden Units)



[Neyshabur, Tomioka, Srebro. ICLR'15]

Experiments with varying number of hidden units





What we learned

- A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks
- PAC-Bayesian theory can partly capture the generalization behavior in deep learning.
- How to use these understanding in practice?

Optimization is Tied to Choice of Geometry

Steepest descent w.r.t. a geometry:

$$w^{(t+1)} = \arg\min_{w} \eta \langle \nabla \mathrm{L}(\mathrm{w}^{(t)}), w \rangle + \delta(w^{(t+1)}, w)$$

✓improve the objective as much as possible

✓only a small change in the model.

Examples:

- Gradient Descent: Steepest descent w.r.t ℓ_2 norm
- Coordinate Descent: Steepest descent w.r.t. ℓ_1 norm
- Path-SGD: Steepest descent w.r.t path-ℓ_2 norm

What's the geometry appropriate for deep networks?


Studying the landscape in search of a flat minimum in Alaska...