Exploring Generalization in Deep Learning

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, Nathan Srebro

Presented by Will Harvey

Contents

- Introduction
- Candidate capacity measures
- Experiments

In previous weeks

- We’ve seen that optimisation algorithms such as SGD provide some implicit regularisation
- Different algorithmic choices (e.g. batch size) lead to minima with different generalisation properties
- Considered sharpness as a measure to explain generalisation performance of different optima, but seen it is flawed

Capacity Measures

"What is the bias introduced by ... algorithmic choices for neural networks? What ensures generalization in neural networks? What is the relevant notion of complexity or capacity control?"

Features of optima, rather than of optimisation algorithms or architectures:
importantly, not uniform across all functions representable by a given architecture.

E.g.
- norm in linear regression
- trace-norm / max-norm in matrix factorisation
- sharpness in deep learning

What properties should a complexity measure have?

Observations:

- our optimization algorithms bias us towards less complex models
- it is possible to capture real data using networks of low complexity

These lead to tests for good complexity measures:

1. We expect the networks learned using real labels to have much lower complexity than those learned using random labels
2. We expect a correlation between the complexity measure and generalisation among zero-training error models.
3. We expect to see the complexity measure decrease as we increase the number of hidden units (and thus improve generalisation).

Notation:

model class: $\mathcal{H}$ *training set*: S *hypothesis*: $h \in \mathcal{H}$

capacity of a model class: number of examples needed to generalise well

complexity measure: $\mathcal{M} : \{\mathcal{H}, S\} \rightarrow \mathbb{R}^+$

restricted model class: $\mathcal{H}_{\mathcal{M},\alpha} = \{h : h \in \mathcal{H}, \mathcal{M}(h) \leq \alpha\}$

Notation: neural networks and losses

Consider *d*-layer feedfoward networks with ReLU activations:

$f_{\mathbf{w}}(\mathbf{x}) = W_d\, D_{d-1}\, W_{d-1}\, \cdots\, D_1\, W_1\, \mathbf{x} = W_d \left(\Pi_{i=1}^{d-1} D_i W_i\right) \mathbf{x}$ is a *d*-layer neural network

where $W_{1,\ldots,d}$ are the weight matrices at each layer and $D_{1,\ldots,d}$ are the ReLU activations (which depend on $\mathbf{x}$). Each hidden layer has dimension h_i.

$\ell(\mathbf{w}, \mathbf{x})$ is the loss on $\mathbf{x}$.

$L(\mathbf{w})$ is the expected loss

$\widehat{L}(\mathbf{w})$ is the empirical loss over the training set

Candidate capacity measures

1. Network size
2. Norms and margins
3. Lipschitz continuity and robustness
4. Sharpness

Network size (parameter count)

Vapnik–Chervonenkis (VC) dimension:

the cardinality of the largest set of points
that the algorithm can shatter

Given a network of depth *d* and number of parameters, **dim(w)**, we can bound the VC dimension as follows:

$$\text{VC-dim} = \tilde{O}(d * \dim(\mathbf{w}))$$

this is not useful when considering networks with more data points than parameters.

Norms and margins

1. ℓ_2 norm
2. ℓ_1-path norm
3. ℓ_2-path norm
4. spectral norm

$$\phi_p(W) = \left(\sum_{j \in \prod_{k=0}^{d} [h_k]} \prod_{i=1}^{d} |W_i[j_i, j_{i-1}]|^p \right)^{\frac{1}{p}}$$

Norms and margins

Problem:

With a 0-1 loss, the scaling of the output is arbitrary. Therefore any norm can be scaled arbitrarily by scaling the output.

Solution:

Quantify the scaling of the output using the *margin* of a data point:

$$\gamma_{\text{margin}}(\mathbf{x}) = f_{\mathbf{w}}(\mathbf{x})[y_{\text{true}}] - \max_{y \neq y_{\text{true}}} f_{\mathbf{w}}(\mathbf{x})[y]$$

And define the margin over a dataset as the smallest γ such that the proportion of data points $\mathbf{x}$ with $\gamma_{\text{margin}}(\mathbf{x}) < \gamma$ is at most ϵ. They use $\epsilon = 0.05$.

Then scale the norms by the margin over the training set.

Norms and margins

Capacity bounds

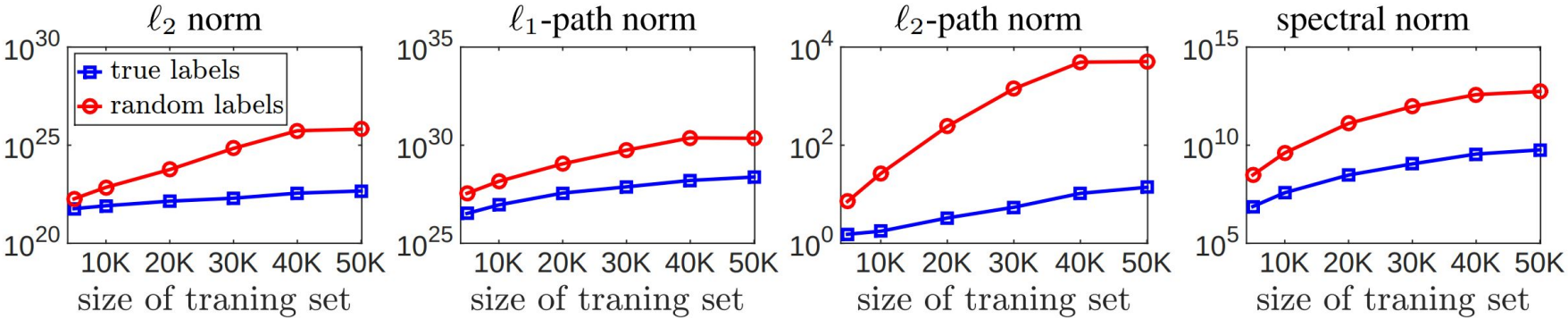
1. ℓ_2 norm : $\frac{1}{\gamma_{\text{margin}}^2} \prod_{i=1}^{d} 4 \|W_i\|_F^2$

2. ℓ_1-path norm : $\frac{1}{\gamma_{\text{margin}}^2} \left| \sum_{j \in \prod_{k=0}^{d}[h_k]} \left| \prod_{i=1}^{d} 2W_i[j_i, j_{i-1}] \right| \right|^2$

3. ℓ_2-path norm : $\frac{1}{\gamma_{\text{margin}}^2} \sum_{j \in \prod_{k=0}^{d}[h_k]} \prod_{i=1}^{d} 4h_i W_i^2[j_i, j_{i-1}]$

4. spectral norm: $\frac{1}{\gamma_{\text{margin}}^2} \prod_{i=1}^{d} h_i \|W_i\|_2^2$

Norms and margins



As expected:

- complexity of models trained on random labels is greater than on real labels
- capacity of model trained on random labels increases faster as the number of labels increases

More experimental validation later on.

Lipschitz continuity and robustness

Given an input space $\mathcal{X}$ and metric $\mathcal{M}$, a function $f : \mathcal{X} \to \mathbb{R}$ on a metric space $(\mathcal{X}, \mathcal{M})$ is called a Lipschitz function if there exists a constant $C_{\mathcal{M}}$, such that $|f(x) - f(y)| \leq C_{\mathcal{M}}\mathcal{M}(x, y)$.

Then the capacity is proportional to $\left(\frac{C_{\mathcal{M}}}{\gamma_{\text{margin}}}\right)^n \text{diam}_{\mathcal{M}}(\mathcal{X})$

where $\text{diam}_{\mathcal{M}}(\mathcal{X}) = \sup_{x,y \in X} \mathcal{M}(x, y)$

Weak bound as it is exponential in input dimension.

Lipschitz continuity and robustness

We can use the L1 path norm ($\prod_{i=1}^{d} \|W_i\|_{1,\infty}$) as a Lipschitz constant.

Then the bound scales as:

$$\left(\frac{\prod_{i=1}^{d} \|W_i\|_2}{\gamma_{\text{margin}}} \right)^n$$

- Exponential in both input dimension and depth

Sharpness

(or vulnerability to adversarial perturbations)

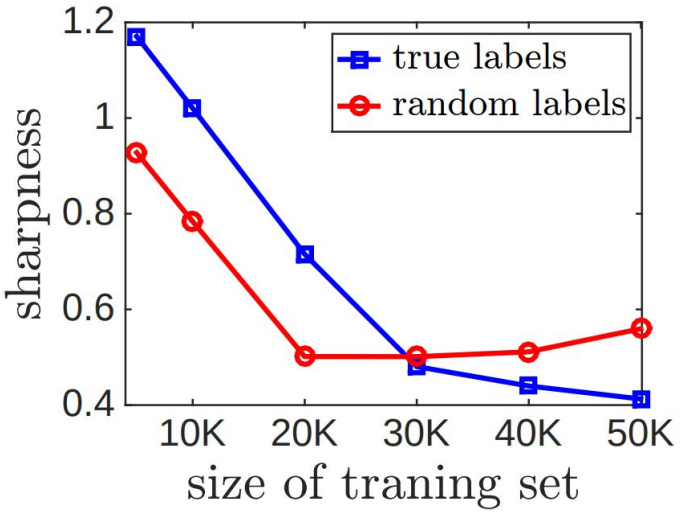
$$\zeta_\alpha(\mathbf{w}) = \frac{\max_{|\boldsymbol{\nu}_i| \leq \alpha(|\mathbf{w}_i|+\mathbf{1})} \widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}}) - \widehat{L}(f_{\mathbf{w}})}{1 + \widehat{L}(f_{\mathbf{w}})} \simeq \max_{|\boldsymbol{\nu}_i| \leq \alpha(|\mathbf{w}_i|+\mathbf{1})} \widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}}) - \widehat{L}(f_{\mathbf{w}}),$$

Sharpness

$$\zeta_\alpha(\mathbf{w}) = \frac{\max_{|\boldsymbol{\nu}_i| \le \alpha(|\mathbf{w}_i|+\mathbf{1})} \widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}}) - \widehat{L}(f_{\mathbf{w}})}{1 + \widehat{L}(f_{\mathbf{w}})} \simeq \max_{|\boldsymbol{\nu}_i| \le \alpha(|\mathbf{w}_i|+\mathbf{1})} \widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}}) - \widehat{L}(f_{\mathbf{w}}),$$

As we saw last week, this fails as a measure when the scale of the parameters is changed.

Additionally, sharpness seems to fail to predict generalisation for networks trained on fewer labels:



Sharpness

PAC-Bayesian perspective

Consider a predictor $f_{\mathbf{w}}$ with parameters $\mathbf{w}$ learnt from the training set. Suppose we have a random variable, $\boldsymbol{\nu}$, and a prior distribution, P, over the hypothesis. Then, with probability $1-\delta$, the following holds:

$$\mathbb{E}_{\boldsymbol{\nu}}[L(f_{\mathbf{w}+\boldsymbol{\nu}})] \leq \mathbb{E}_{\boldsymbol{\nu}}[\widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}})] + 4\sqrt{\frac{\left(KL\left(\mathbf{w}+\boldsymbol{\nu}\|P\right) + \ln\frac{2m}{\delta}\right)}{m}}$$

Valid for any prior and perturbation distribution.

Sharpness

PAC-Bayesian perspective

$$\mathbb{E}_{\boldsymbol{\nu}}[L(f_{\mathbf{w}+\boldsymbol{\nu}})] \leq \mathbb{E}_{\boldsymbol{\nu}}[\widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}})] + 4\sqrt{\frac{\left(KL\left(\mathbf{w}+\boldsymbol{\nu}\|P\right) + \ln\frac{2m}{\delta}\right)}{m}}$$

$$\mathbb{E}_{\boldsymbol{\nu}}[L(f_{\mathbf{w}+\boldsymbol{\nu}})] \leq \widehat{L}(f_{\mathbf{w}}) + \underbrace{\mathbb{E}_{\boldsymbol{\nu}}[\widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}})] - \widehat{L}(f_{\mathbf{w}})}_{\text{expected sharpness}} + 4\sqrt{\frac{1}{m}\left(KL\left(\mathbf{w}+\boldsymbol{\nu}\|P\right) + \ln\frac{2m}{\delta}\right)}$$

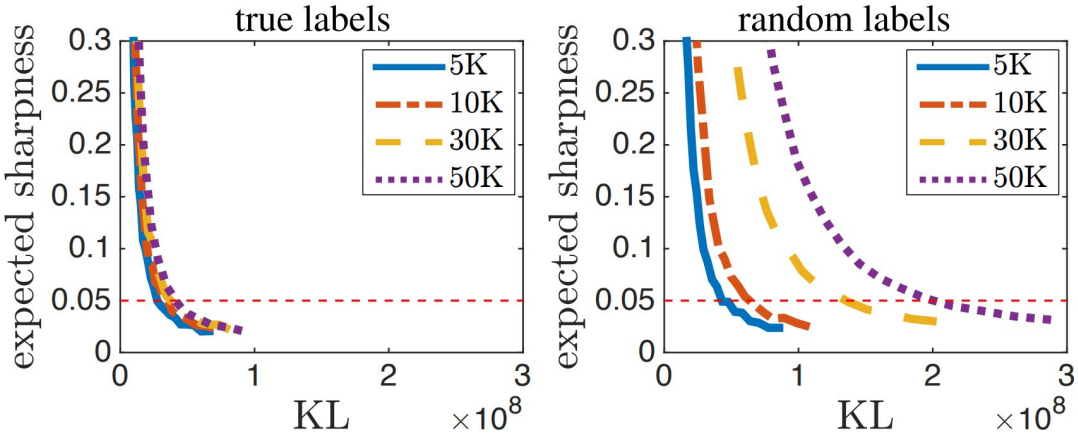
Then we set the prior and perturbation distributions to be spherical Gaussians:

$$\mathbb{E}_{\boldsymbol{\nu}\sim\mathcal{N}(0,\sigma)^n}[L(f_{\mathbf{w}+\boldsymbol{\nu}})] \leq \widehat{L}(f_{\mathbf{w}}) + \underbrace{\mathbb{E}_{\boldsymbol{\nu}\sim\mathcal{N}(0,\sigma)^n}[\widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}})] - \widehat{L}(f_{\mathbf{w}})}_{\text{expected sharpness}} + 4\sqrt{\frac{1}{m}\left(\underbrace{\frac{\|\mathbf{w}\|_2^2}{2\sigma^2}}_{\text{KL}} + \ln\frac{2m}{\delta}\right)},$$

Sharpness

PAC-Bayesian perspective

$$\mathbb{E}_{\boldsymbol{\nu}\sim\mathcal{N}(0,\sigma)^n}\left[L(f_{\mathbf{w}+\boldsymbol{\nu}})\right] \leq \widehat{L}(f_{\mathbf{w}}) + \underbrace{\mathbb{E}_{\boldsymbol{\nu}\sim\mathcal{N}(0,\sigma)^n}\left[\widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}})\right] - \widehat{L}(f_{\mathbf{w}})}_{\text{expected sharpness}} + 4\sqrt{\frac{1}{m}\left(\underbrace{\frac{\|\mathbf{w}\|_2^2}{2\sigma^2}}_{\text{KL}} + \ln\frac{2m}{\delta}\right)},$$



The combination of KL and expected sharpness seems to behave sensibly.

Experiments

Experiments

Back to our three tests:

1. Networks learned using real labels should have much lower complexity than those learned using random labels.
2. We expect a correlation between the complexity measure and generalisation among zero-training error models.
3. We expect to see the complexity measure decrease as we increase the number of hidden units.

Experiments

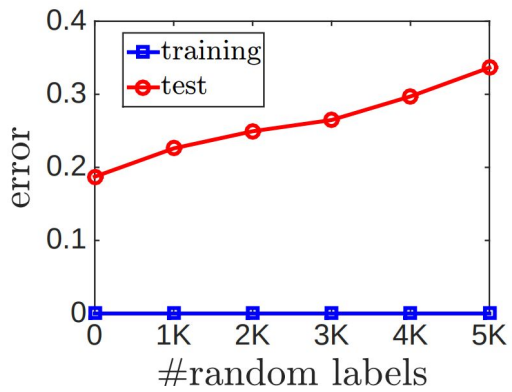
We have seen results from training on real vs. random data. They consider two more experiments:

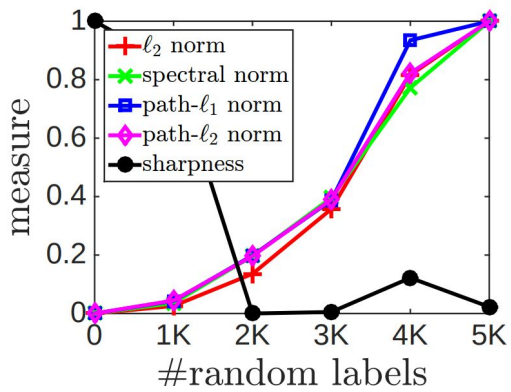
1. Training on the union of 10 000 CIFAR10 training images and a randomly labelled confusion set. This should lead to networks with zero training error that generalise poorly. As the size of the confusion set increases, model capacity should increase as test performance decreases.

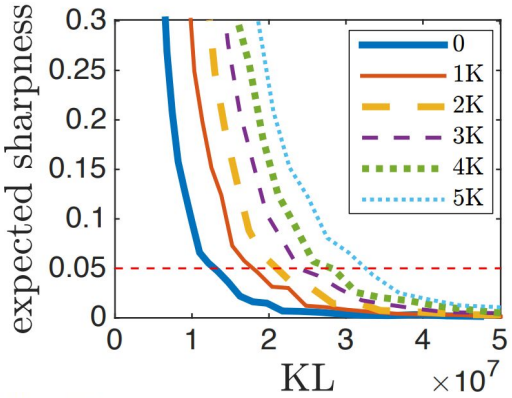
2. Increasing the network size.They train a fully-connected, single-hidden layer MNIST classifier with a varying hidden layer size. All achieve zero-training error, but larger hidden layers achieve better generalisation. This should be reflected in a good capacity measure.

Further Experiments

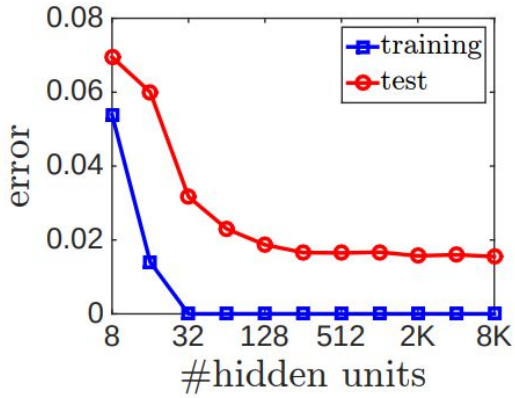
1.

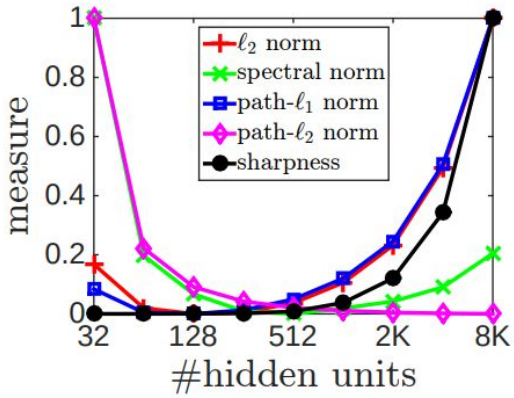


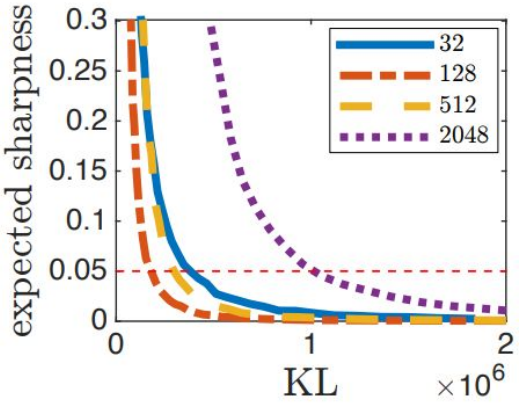




2.







New Generalisation Bound

They develop a bound by considering conditions to keep sharpness low:

1. Prevent weak interactions between neighbouring layers.
2. Prevent small perturbations in weights causing large changes in the number of activations.
3. Prevent nodes in lower layers with large weight becoming active, causing potentially large changes in output

$(C1):$ Given x, let $x = W_0$ and $D_0 = I$. Then, for all $0 \leq a < c < b \leq d$, $\| \left(\Pi_{i=a}^{b} D_i W_i\right) \|_F \geq \frac{\mu}{\sqrt{h_c}} \|\Pi_{i=c+1}^{b} D_i W_i\|_F \| \left(\Pi_{i=a}^{c} D_i W_i\right) \|_F$.

$(C2):$ Given x, for any level k, $\frac{1}{h_k} \sum_{i \in [h_k]} 1_{W_{k,i} \Pi_{j=1}^{k-1} D_j W_j x \leq \delta} \leq C_2 \delta$.

$(C3):$ For all i, $\|W_i\|_{2,\infty}^2 h_i \leq C_3^2 \|D_i W_i\|_F^2$.

$$\mathbb{E} \left| \widehat{L}(f_{\mathbf{w}+\boldsymbol{\nu}}(x)) - \widehat{L}(f_{\mathbf{w}}(x)) \right|$$
$$\leq \left[\Pi_{i=1}^{d} (1+\gamma_i) - 1 + \Pi_{i=1}^{d} (1+\gamma_i C_2 C_3) \left(\Pi_{i=1}^{d} (1+\gamma_i C_\delta C_2) - 1 \right) \right] C_L \|f_{\mathbf{w}}(x)\|_F .$$

Conclusion

- Proposed three tests for the usefulness of a capacity measure.
- Reviewed pre-existing measures of capacity.
- Developed a new measure based on PAC-Bayes.