Generalization and Learning Theory for Deep Learning

Jacob Seidman

University of Pennsylvania

seidj@sas.upenn.edu

October 11, 2018

Outline

- Classical Learning Theory Bounds
- Expressiveness of Neural Networks
- Norms, Margins, and Sharpness
- PAC-Bayes results
- Compression
- Some unanswered questions

Standard Generalization Bound

- What does a generalization guarantee look like?
- (Binary Classification) Let P be the true distribution over $\mathcal{X} \times \{\pm 1\}$ and $\hat{P}_n$ the empirical distribution from n samples of P, $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
- $f : \mathcal{X} \to \{\pm 1\}$ coming from a class F. With high probability over the sample,

$$P(Y \neq f(X)) \leq \hat{P}_n(Y \neq f(X)) + \frac{c}{\sqrt{n}}\text{complexity}(F)$$

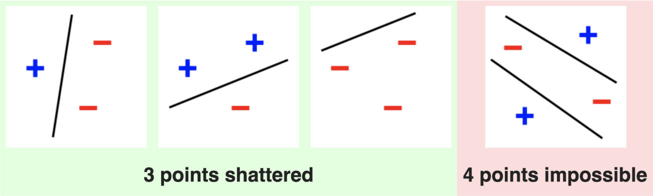
- For a fixed $f \in F$, have a concentration inequality

$$\mathbb{P}(\text{difference of empirical and real loss} \leq \epsilon) \leq 1 - e^{-\epsilon^2 m}.$$

- Do a "union bound" over all possible $f \in F$.
- What measure of complexity should we use for our "union bounding"?

VC-dimension

- A function class F *shatters* a set of k points $(x_1, \ldots, x_n)$ if for any assignment of labels $\{\pm 1\}$ to the x_i, there exists $f \in F$ such that f gives the desired assignment of labels.
- The VC-dimension of a class of functions F is the cardinality of the largest set of points that can be shattered by F.
- The VCdim is agnostic to any structure of the distribution P we sample from.



Theorem

(Vapnik, Chernovenkis, 1971) Let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a sample drawn i.i.d from P and $\delta > 0$. Then there exists a constant C such that for any n and all $f \in F$, with probability $1 - \delta$,

$$P(Y \neq f(X)) \leq \hat{P}_n(Y \neq f(X)) + c\sqrt{\frac{\mathit{VCdim}(F)}{n}}$$

- A d-layer ReLU activated neural network has up to logarithmic factors VC-dim $\tilde{O}(d \cdot (\# \text{ of parameters}))$.

Distributionally Dependent Complexities

- Given a distribution μ over $\mathcal{X}$, i.i.d samples $X_1, \dots, X_n$, and a class of functions F from $\mathcal{X}$ to $\mathbb{R}$:

Definition

The maximum discrepancy of F is the random variable

$$\hat{D}_n(F) = \sup_{f \in F} \left(\frac{2}{n} \sum_{i=1}^{n/2} f(X_i) - \frac{2}{n} \sum_{i=n/2+1}^{n} f(X_i) \right),$$

the expected maximum discrepancy is

$$D_n(F) = \mathbb{E}_\mu[\hat{D}_n(F)]$$

Definition

Let $\sigma_1, \ldots, \sigma_n$ be independent Bernoulli(1/2) random variables. Define

$$\hat{R}_n(F) = \mathbb{E}_\sigma \left[\sup_{f \in F} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right| \mid X_1, \ldots, X_n \right].$$

The *Rademacher complexity* is

$$R_n(F) := \mathbb{E}_\mu[\hat{R}_n(F)].$$

Definition

Let $g_1, \ldots, g_n$ be independent Gaussian(0,1) random variables. Define

$$\hat{G}_n(F) = \mathbb{E}_g \left[\sup_{f \in F} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right| \mid X_1, \ldots, X_n \right].$$

The *Gaussian complexity* is

$$G_n(F) := \mathbb{E}_\mu[\hat{G}_n(F)].$$

• We have the following relations for some constants c and C,

$$cR_n(F) \le G_n(F) \le C \log n R_n(F),$$

and if F is a class of functions mapping into $[-1, 1]$,

$$\frac{R_n(F)}{2} - 2\sqrt{\frac{2}{n}} \le D_n(F) \le R_n(F) + 4\sqrt{\frac{2}{n}}.$$

• Rademacher and Gaussian complexities take the data generating distribution into account.

- Quantify how much a function from f can be correlated with a noise sequence of length n.

A PAC Theorem

Theorem

(Bartlett, Mendelson 2002) F *is a set of* $\{\pm 1\}$ *valued functions defined on* $\mathcal{X}$. $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ *is a sequence of i.i.d samples from* P. *With probability at least* $1 - \delta$, *for every* $f \in F$,

$$P(Y \neq f(X)) \leq \hat{P}_n(Y \neq f(X)) + \hat{D}_n(F) + \sqrt{\frac{9 \log(1/\delta)}{2n}}$$

and

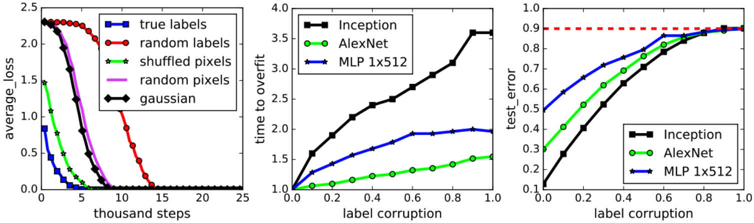
$$P(Y \neq f(X)) \leq \hat{P}_n(Y \neq f(X)) + \frac{R_n(F)}{2} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

An Experiment

- Are these bounds useful?

- We can get a sense of the Rademacher complexity for neural networks in a classification problem by trying to fit random data.

- (Zhang et. al. ICLR 2017) did this: Trained neural networks (Inception V3, Alexnet, MLP) on versions of CIFAR10 and ImageNet with
  - True labels
  - Partially corrupted labels (with probability p each label is changed to a uniformly random label)
  - Random labels
  - Randomly permuted pixels (same permutation across all images)
  - Independently chosen random permutation for each image
  - Gaussian pixels instead of each image



(a) learning curves (b) convergence slowdown (c) generalization error growth

Theorem

(Zhang et. al. ICLR 2017) There exists a two-layer neural network with ReLU activations and $2n + d$ parameters that can represent any function on a sample of n points in d dimensions.

- Proof by expressing fitting problem as a full rank linear system.
- Can be extended to depth k network with width $O(n/k)$.

Generalization in a Linear Model

- Consider fitting a linear model to $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$ feature vectors, $y_i \in \mathbb{R}$, $d > n$
- Let X be the data matrix such that the ith row of X is $x_i^\top$. Fitting the linear model is solving the system $Xw = y$.
  - If $\text{rank}(X) = n$ then there are infinitely many solutions.
  - Which one generalizes best?

Generalization in a Linear Model with SGD

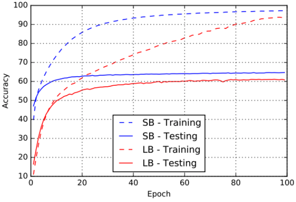
- Use some convex loss $\ell(x_i, y_i)$ and train with classic SGD (sample one point to compute gradient at each iterate).
- If initial iterate is $w_0 = 0$:
  - SGD converges to some $w \in \text{span}\{x_1, \ldots, x_n\}$; for some $\alpha \in \mathbb{R}^n$, $w = X^\top \alpha$.
  - If the training error is 0, then $Xw = y$
- Previous two points imply that $XX^\top \alpha = y$. This linear system has a unique solution!
  - This also turns out to be the minimum norm solution of the original problem.
- This model actually works on CIFAR and MNIST with some preprocessing and enough memory

Recap

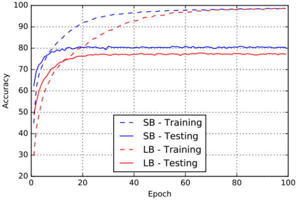
- So far:
  - VC-dimension, Rademacher/Gaussian Complexity are not useful capacity/complexity measures for explaining generalization.
  - Norms seem to be somewhat useful but don't explain the whole story.
  - Choices of what kind of norms to use.
- Can we characterize generalization ability by the nature of what local minimum we converge to?
- If U is a neighborhood of a minimum x, define the *sharpness* of a local minimum as

$$\frac{\max_{y \in U} \quad f(y) - f(x)}{f(x) + 1}.$$

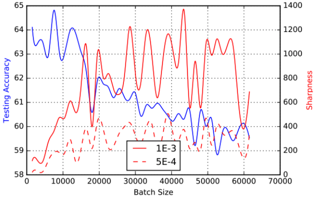
An Experiment (Keskar et. al. ICML 2017)



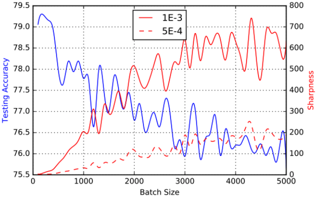
(a) Network F_2



(b) Network C_1



(a) F_2



(b) C_1

Batch sizes and Noise

- Continuous limit of SGD can be written as (Li et. al. ICML 2017)

$$dw(t) = -\nabla_{\mathbf{w}} \hat{L}(f_{\mathbf{w}})\, dt + \sqrt{2\beta^{-1} D(\mathbf{w})} dW(t),$$

where $\beta^{-1} \propto$ (step size)/(batch size)

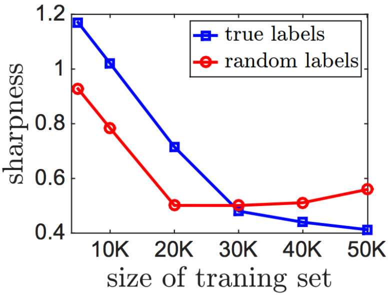
- Interpretation: A proper amount of noise in the dynamics makes it more likely for the algorithm to stay away from sharp minima.

Sharpness?

- Is sharpness the kind of measure we are looking for?

- (Dinh et. al. ICML 2017) Show that by reparameterizing the function arbitrarily sharp minima can be created that have the same generalization ability.

- (Neyshabur et. al. NIPS 2017)



- Sharpness alone is not enough to explain generalization.

A PAC Bayes Framework

- Denote a d-layer feedforward neural network with parameter vector $\mathbf{w} \in \Omega$ as $f_{\mathbf{w}}(x) = W_d\phi(W_{d-1}\phi(\ldots\phi(W_1 x)))$, where ϕ is a nonlinear activation.

- Let $L(f_{\mathbf{w}})$ and $\hat{L}(f_{\mathbf{w}})$ be the true loss and emprical loss, respectively, for the neural network $f_{\mathbf{w}}$.

Theorem

(McAllester 2003) Given a prior distribution over the parameter space Q_0, independent of the training data, for any $\delta \in (0,1)$ and any random variable $\boldsymbol{\nu}$, the following holds with probability $1-\delta$,

$$\mathbb{E}_{\boldsymbol{\nu}}[L(f_{\mathbf{w}+\boldsymbol{\nu}})] \leq \mathbb{E}_{\boldsymbol{\nu}}[\hat{L}(f_{\mathbf{w}+\boldsymbol{\nu}})] + 4\sqrt{\frac{1}{n}\left(\mathit{KL}(\mathbf{w}+\boldsymbol{\nu}||Q_0) + \log\frac{2n}{\delta}\right)}.$$

- If we take $\boldsymbol{\nu}$ such that $\mathbb{E}[\boldsymbol{\nu}] = 0$ and $\boldsymbol{\nu}$ is concentrated in some small neighborhood of 0, then from the previous theorem

$$\mathbb{E}_{\boldsymbol{\nu}}[L(f_{\mathbf{w}+\boldsymbol{\nu}})] \leq \hat{L}(f_{\mathbf{w}}) + \underbrace{\mathbb{E}_{\boldsymbol{\nu}}[\hat{L}(f_{\mathbf{w}+\boldsymbol{\nu}})] - \hat{L}(f_{\mathbf{w}})}_{\text{expected sharpness}} + 4\sqrt{\frac{1}{n}\left(\mathsf{KL}(\mathbf{w}+\boldsymbol{\nu}||Q_0) + \log\frac{2n}{\delta}\right)}.$$

- Generalization controlled by sharpness *and* distance away from prior.

To give a more specific example:

- Let P and $\boldsymbol{\nu}$ be independent 0 mean isotropic gaussians with variance σ^2.

$$\mathbb{E}_{\boldsymbol{\nu}}[L(f_{\mathbf{w}+\boldsymbol{\nu}})] \leq \hat{L}(f_{\mathbf{w}}) + \underbrace{\mathbb{E}_{\boldsymbol{\nu}}[\hat{L}(f_{\mathbf{w}+\boldsymbol{\nu}})] - \hat{L}(f_{\mathbf{w}})}_{\text{expected sharpness}} + 4\sqrt{\frac{1}{n}\left(\frac{\|\mathbf{w}\|_2^2}{2\sigma^2} + \log\frac{2n}{\delta}\right)}.$$

- Can we do something similar with another kind of norm?

Margin Loss

- For a distribution $\mathcal{D}$ and classifier f define the *margin loss* as

$$L_\gamma(f_\mathbf{w}) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[f_\mathbf{w}(x)[y] \leq \gamma + \max_{j\neq y} f_\mathbf{w}(x)[j]\right]$$

- Let $\hat{L}_\gamma(f_\mathbf{w})$ be the empirical margin loss.
- Note: $L_0(f_\mathbf{w}) = L(f_\mathbf{w})$ and $\hat{L}_0(f_\mathbf{w}) = L(f_\mathbf{w})$.

(Neyshabur et. al. ICLR 2018)

- Fix prior P independent of the data, γ and take a perturbation $\boldsymbol{\nu}$ such that

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

Then

$$L_0(f_{\mathbf{w}}) \leq \hat{L}_\gamma(f_{\mathbf{w}}) + 4\sqrt{\frac{\mathsf{KL}(w+\boldsymbol{\nu}||P) + \log\frac{6n}{\delta}}{n-1}}.$$

• Let $\mathcal{X}_{B,m}$ be the ball of radius B centered at the origin in $\mathbb{R}^m$. For any $\mathbf{w} \in \mathcal{X}_{B,m}$ and any perturbation vector $\boldsymbol{\nu} = \text{vec}(\{U_i\}_{i=1}^d)$ such that $\|U_i\|_2 \leq \frac{1}{d}\|W_i\|_2$,

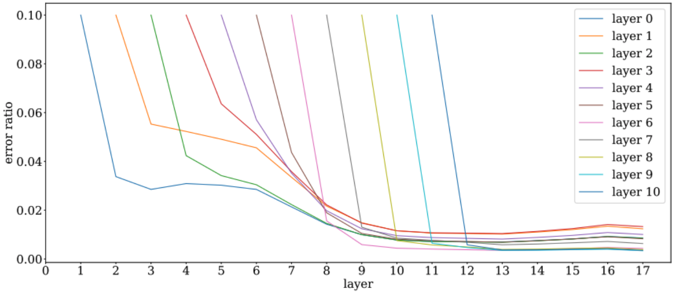
$$|f_{\mathbf{w}+\boldsymbol{\nu}}(x) - f_{\mathbf{w}}(x)|_2 \leq eB\left(\prod_{i=1}^d \|W_i\|_2\right)\sum_{i=1}^d \frac{\|U_i\|_2}{\|W_i\|_2}.$$

• This bounds a measure of the sharpness in terms of the spectral norms of the layers.

- A generalization bound: For any $\delta, \gamma > 0$, with probability at least $1 - \delta$ we have

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

- The right hand side is interesting but too large.



- Gaussian noise injected with input into trained NN (CIFAR-10). Error ratio is relative difference in activations for each layer.
- Suggests we can compress the network.

A notion of Compressibility

- Let f be a classifier and $G_\mathcal{A} = \{g_A \mid A \in \mathcal{A}\}$ be a set of classifiers. f is (γ, S)-compressible via $G_\mathcal{A}$ if there exists $A \in \mathcal{A}$ such that for any $x \in S$, we have for all y,

$$|f(x)[y] - g_A(x)[y]| \leq \gamma.$$

- Let $G_{\mathcal{A},s} = \{g_{\mathcal{A},s} \mid A \in \mathcal{A}\}$ be a set of classifiers indexed by a helper strings s. f is (γ, S) compressible with respect to $G_{\mathcal{A},s}$ using helper string s if there exists $A \in \mathcal{A}$ such that for any $x \in S$, we have for all y,

$$|f(x)[y] - g_{A,s}(x)[y]| \leq \gamma.$$

Theorem

(Arora et. al. ICML 2018) Let $G_{\mathcal{A},s} = \{g_{\mathcal{A},s} \mid A \in \mathcal{A}\}$ *be a set of classifiers, where* A *is a set of* q *parameters, each of which can take at most* r *values and* s *is a helper string. If* f *is* (γ, S) *compressible via* $G_{\mathcal{A},s}$*, with* S *being a training sample of* n *examples, then there exists* $A \in \mathcal{A}$ *such that with high probability*

$$L_0(g_{A,s}) \leq \hat{L}_\gamma(f) + O\left(\sqrt{\frac{q \log r}{m}}\right).$$

- This can recover the theorem from Neyshabur et. al. ICLR 2018.

- We will need some definitions to get our compressibility and therefore generalization guarantee for neural networks.

- If $M : \mathbb{R}^d \to \mathbb{R}^\ell$ and $\mathcal{N}$ is some noise distribution, then the *noise sensitivity* of M at x with respect to $\mathcal{N}$ is

$$\psi_{\mathcal{N}}(M, x) = \mathbb{E}_{\eta \sim \mathcal{N}} \left[\frac{\|M(x + \eta\|x\|) - M(x)\|^2}{\|M(x)\|^2} \right].$$

- If $\mathcal{N}$ is a mean 0 unit Gaussian distribution then

$$\psi_{\mathcal{N}}(M, x) = \frac{\|M\|_F^2 \|x\|^2}{\|Mx\|^2}.$$

More definitions

- The *layer cushion* of layer i is the largest number μ_i such that for all $x \in S$,

$$\mu_i \|W_i\|_F \|\phi(x^{i-1})\| \leq \|W_i \phi(x^{i-1})\|.$$

•Let $M^{i,j}$ be the operator from the ith layer of the network to the jth, and $J^{i,j}$ its Jacobian.

- For $i \leq j$, the *interlayer cushion* $\mu_{i,j}$ is the largest number such that for any $x \in S$,

$$\mu_{i,j} \|J_x^{i,j}\|_f \|x\| \leq \|J_x^{i,j} x\|$$

For a layer i the *minimal interlayer cushion* is

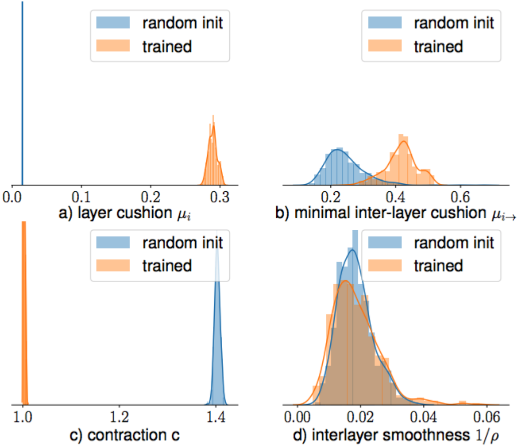
$$\mu_{i\rightarrow} := \text{min}_{i \leq j \leq d} \mu_{i,j}.$$

- The *activation contraction* is the smallest number c such that for any layer i and any $x \in S$,

$$\|\phi(x)\| \geq \frac{\|x\|}{c}.$$

Noise sensitivity measures

- How do these measures of noise sensitivity change over training?



A Compression Generalization Theorem

Theorem

(Arora et. al. ICML 2018) (Informal) If for a fully connected network f_W, $(W = \{W_1, \dots, W_d\})$ we can project the weight matrices onto a random set of sensing matrices such that the effective noise introduced is passed nearly linearly through the layers, then for any $\delta \in (0,1)$ we have that with probability $1-\delta$, for any $\gamma > 0$, the compressed version of f_W with weight matrices $\tilde{W}$ satisfies,

$$L_0(f_{\tilde{W}}) \leq \hat{L}_\gamma(f_{\tilde{W}}) + \tilde{O}\left(\sqrt{\frac{c^2 d^2 \max_{x \in S} \|f_A(x)\|_2^2 \sum_{i=1}^d \frac{1}{\mu_i^2 \mu_{i\rightarrow}^2}}{\gamma^2 m}}\right).$$

Algorithm 1 Matrix-Project $(A,\ \varepsilon,\ \eta)$

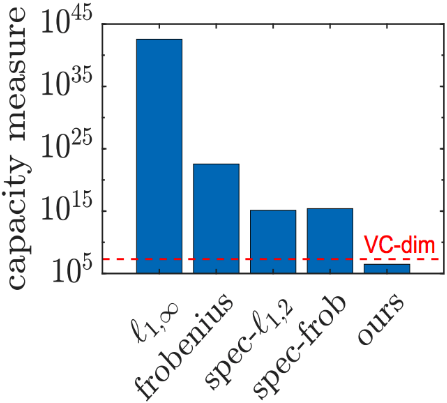
Require: Layer matrix $A \in \mathbb{R}^{h_1 \times h_2}$, error parameter ε, η.
Ensure: Returns $\hat{A}$ s.t. $\forall$ fixed vectors u, v,

$$\Pr[\|u^\top \hat{A} v - u^\top A v\| \geq \varepsilon \|A\|_F \|u\| \|v\|] \leq \eta.$$

Sample $k = \log(1/\eta)/\varepsilon^2$ random matrices $M_1, \ldots, M_k$ with entries i.i.d. ± 1 ("helper string")
for $k' = 1$ to k **do**
  Let $Z_{k'} = \langle A, M_{k'} \rangle M_{k'}$.
end for
Let $\hat{A} = \frac{1}{k} \sum_{k'=1}^{k} Z_{k'}$

Performance

- So did Arora et. al. (2018) finally find a useful bound?



- Closer.

Further Questions

- Proof of compressibility properties of neural networks
- Dependence on structure of training data?
- How to define structure of training data?
- Implicit/explicit regularization from training methods
  - Are we actually being pushed toward a smaller/less complex function space?
- Structure of (implicit/surrogate) loss landscape

Thank you!