Mathematics of Deep Learning
Lecture 10: Mean Field Regime, Rademacher Complexity, and NNs Generalization Bounds

Prof. Predrag R. Jelenković
Time: Tuesday 4:10-6:40pm

Dept. of Electrical Engineering
Columbia University , NY 10027, USA
Office: 812 Schapiro Research Bldg.
Phone: (212) 854-8174
Email: predrag@ee.columbia.edu
URL: http://www.ee.columbia.edu/∼predrag

Overparametrization: Lazy and Mean Field Regimes

Lazy regime: weights don't move much. The following paper pinpoints the key ingredient responsible for wights remaining close to the initial values during training

- A Note on Lazy Training in Supervised Differentiable Programming, by L. Chizat and F. Bach, Dec 2018.

  See also an updated version:
  On Lazy Training Differentiable Programming, by L. Chizat and F. Bach, NIPS 2019.

- Chapter 8 in the recent monograph Deep Learning Theory Lecture Notes, by Telgarsky, Oct 2021.

Mean Filed regime: weight move a lot

- On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport, by L. Chizat and F. Bach, NIPS 2018.
- A Mean Field View of the Landscape of Two-Layers Neural Networks, Mei, Song, Andrea Montanari, and Phan-Minh Nguyen, 2018.

Lazy vs Active: Many Neurons Example

- Ground truth: x uniform on a unit sphere, y output of NN with 3 neurons
- $n = 200$ data samples
- Model wide network: $m = 200$ neurons
- τ - variance of $w_{ij}(0)$

(a) "Active" training ($\tau = 0.1$) (b) Lazy ($\tau = 2$)

Hence, the key to NTK regime is the scaling, rather than over-parametrization.

Recall Tangent Model

Consider any parametric model $f(w,x), w \in \mathbb{R}^p, x \in \mathbb{R}^d$. Then, assuming f is differentiable, we can linearize this model around the initial parameters w_0 as:

$$f_0(w,x) := f(w_0,x) + (w - w_0) \cdot \nabla_w f(w_0,x), \quad \text{(Tangent model)}.$$

If $f(w_0,x) = 0$ (doubling trick), or $f(w_0,x) \approx 0$ (whide network), then

$$f_0(w,x) = (w - w_0) \cdot \nabla_w f(w_0,x),$$

and training this model is equivalent to linear (Kernel) model with features contained in $\nabla_w f(w_0,x)$ Hence, with the quadratic loss,

training $f_0(w,x)$ is equivalent to solving a dual problem in the dot-product space according to neuro-tangent-kernel (NTK):

$$K(x,y) = \nabla_w f(w_0,x) \cdot \nabla_w f(w_0,y).$$

Question: When is training $f_0(w,x)$ going to produce approximately the same results as $f(w,x)$? (Obviously, if $f(w,x)$ is linear, then $f_0(w,x) = f(w,x)$.)

For Wide Networks: Lazy vs Active Training

Hence, in wide networks we can consider two interesting regimes

$$\alpha(m) = \frac{1}{\sqrt{m}} \Rightarrow \mathbb{E}[\kappa_f(\boldsymbol{w}_0)] = O(m^{-1/2}) \ll 1, \quad \text{(NTK/lazy regime)}$$
$$\alpha(m) = \frac{1}{m} \Rightarrow \mathbb{E}[\kappa_f(\boldsymbol{w}_0)] = O(1), \quad \text{(Mean field regime)}$$

In mean field regime

- The weights $\boldsymbol{w}$ move considerably during training
- Much harder problem
- We'll consider in the future some papers in this regime

Result On Lazy Training With General α-Scaling

- Consider n data points $(x_i, y_i), 1 \le i \le n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, and let

$$f(w) = (f(x_1; w), \dots, f(x_n; w))^\top \in \mathbb{R}^n.$$

- Assume quadratic loss

$$L(\alpha f(w)) := \frac{1}{2}\|\alpha f(w) - y\|^2, \quad \alpha > 0.$$

- Then, the gradient flow evolves as

$$\dot{w}(t) = -\nabla_w L(\alpha f(w(t))) = -\alpha J_t^\top \nabla L(\alpha f(w(t))),$$

  where $J_t = Df(w(t))$ is the Jacobian

$$J_t = J_{w(t)} = (\nabla f(x_1; w(t)), \dots, \nabla f(x_1; w(t)))^\top \in \mathbb{R}^{n \times p}$$

- Linear tangent model flow, with the same initialization $w(0)$

$$f_0(u) = f(w(0)) + J_0(u - w(0))$$
$$\dot{u}(t) = \nabla_w L(\alpha f_0(u(t))) = -\alpha J_0^\top \nabla L(\alpha f_0(u(t)))$$

Overparametrized NTK/Lazy Regime

- How close is the nonlinear gradient flow to the linear one?
- Assumptions

$$\text{rank}(J_0) = n$$

$$\sigma_{\min} = \sigma_{\min}(J_0) = \sqrt{\lambda_{\min}(J_0 J_0^\top)} > 0$$

$$\|J_w - J_v\| \leq \beta \|w - v\|$$

$$\alpha \geq \frac{\beta \sigma_{\max} \sqrt{1152 L_0}}{\sigma_{\min}^3}, \quad L_0 = \frac{1}{2}\|\alpha f(w(0)) - y\|^2$$

Theorem 8.1 (Telgarsky, also Theorem 3.2 in Chizat&Bach, 2019)
Under the preceding assumptions,

$$\max(L(\alpha f(w(t))), L(\alpha f_0(u(t)))) \leq L_0 \exp(-t\alpha^2 \sigma_{\min}^2/2)$$

$$\max(\|w(t) - w(0)\|, \|u(t) - w(0)\|) \leq \frac{3\sigma_{\max}\sqrt{8L_0}}{\alpha \sigma_{\min}^2}$$

Simplifying the Result With Ballpark Estimates of Constants

- Smoothness Constant: $\beta = \Theta(n)$
- Singular values: $\sigma_{\min}, \sigma_{\max}$ should scale as $\sqrt{m}$
- $m > n^3$
- Initial Risk: $L_0 = \frac{1}{2}\sum_{i=1}^{n}(\alpha f(x_i) - y_i)^2 = \Theta(\alpha^2 mn)$
- Combining all parameters for $\alpha = 1/\sqrt{m}$, simplifies Theorem 8.1:

$$\max(L(\alpha f(w(t))), L(\alpha f_0(u(t)))) = O\left(n \exp(-\Omega(t))\right)$$

$$\max\left(\frac{1}{\sqrt{m}}\|w(t) - w(0)\|, \frac{1}{\sqrt{m}}\|u(t) - w(0)\|\right) = O\left(\frac{\sqrt{n}}{\sqrt{m}}\right) = o(1),$$

sinnce $m \gg n$.

Mean Field Scaling

- Scaling $\alpha = 1/m$, i.e., the model for one hidden layer is

$$\begin{aligned}\hat{f}(x) &= \alpha(m) \sum_{j=1}^{m} a_j \sigma(w_j \cdot x + b_j) \\ &=: \frac{1}{m} \sum_{j=1}^{m} \phi(\theta_j, x) \\ &\to f(x, \mu) := \mathbb{E}_\theta[\phi(\theta, x)] = \int \phi(\theta, x) d\mu(\theta), \quad \mu \in \mathcal{P}(\mathbb{R}^{d+2}),\end{aligned}$$

  where $\mathcal{P}$ is the space of probability measures.

- Hence, in the limit, when the number of data points is large, we can consider the population loss

$$L(\mu) = \mathbb{E}_{(x,y)}[\ell(f(\mu, x), y)]$$

  over the space of probability measures $\mu \in \mathcal{P}(\mathbb{R}^{d+2})$

Mean Field Scaling Analysis

- Based on Transportation Theory and Wasserstein metric, which measures the distance between the probability measures.
- Uses Wasserstein Gradient Flow:

$$\frac{d}{dt}\mu_t = -\mathsf{div}(\mu_t v_t), \quad \mu_0 \in \mathcal{P}(\mathbb{R}^{d+2})$$

  and $v_t(\theta) = -\nabla(\delta L(\mu_t)/\delta\mu)$, where $\delta L(\mu_t)/\delta\mu$ is the first variational derivative, for a given functional L.
- The analysis is quite involved: more details can be found in Chizat&Bach, NIPS 20018, as well as Mei, Montanari, Nguyen, 2018. In general, they show, under some conditions
  - Gradient flow over empirical measures is close to Wasserstein Gradient Flow:

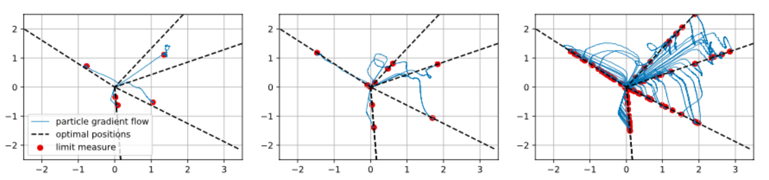
$$\hat{\mu}_t^{(m)} \approx \mu_t$$

  - Empirical measures achieve global population loss minimum

$$\lim_{t,m\to\infty} L(\hat{\mu}_t^{(m)}) = \min_{\mu} L(\mu)$$

Mean Field Examples

- Examples from L. Chizat and F. Bach, NIPS 2018.
- Ground truth: x uniform on a unit circle in $\mathbb{R}^2$, y output of NN with $m = 4$ neurons
- Training a neural network with ReLU activation, and 5, 10, and 100 neurons.



Hence, the key to NTK regime is the scaling, rather than over-parametrization.

Comparison of Mean Field and Lazy Scaling

		Mean Field	Lazy
Model	$f(w,x) =$	$\frac{1}{m}\sum \phi(w_i,x)$	$\frac{1}{\sqrt{m}}\sum \phi(w_i,x)$
Initial predictor	$\|f(w_0,x)\| =$	$O(1/\sqrt{m})$	$O(1)$
Displacement	$\|w_\infty - w_0\| =$	$O(1)$	$O(1/\sqrt{m})$
Relative scale	$\mathbb{E}[\kappa_f(\boldsymbol{w}_0)] =$	$O(1)$	$O(1/\sqrt{m})$

Note:

- Deep NNs are commonly initialized with $\mathrm{Var}(w_{ij}) = O(\sqrt{2/\mathsf{fan}_{\mathsf{in}}})$, which corresponds to the Lazy/linear regime.
- Intuitively, the success of NNs should be due to the nonlinear regime, i.e., the ability to find the best basis, for which ws need to move.

Bounding Generalization Error

The main objective of statistical learning is to predict well on future data.

- How do we measure the error?
  - Regression: Quadratic error/loss

$$\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$$

  - Classification: $\ell(\hat{y}, y) = 1_{\{\hat{y} \neq y\}}$
- $\{\mathcal{H}\}$: Hypothesis class of functions, e.g., all functions, $h(x, w)$, that can be generated by a NN of a certain architecture.
- Empirical Risk/Loss - total training error: For a data sample $S = \{(x_i, y_i)\}_{i=1}^n$ and $h \in \mathcal{H}$

$$\hat{L}_n = \hat{L}_n(h) \equiv L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$$

Shai's book [ML2014] uses $L_S(h)$ notation; we'll use these interchangeably.

Empirical Risk Minimization and True Error

During training one typically minimizes the empirical risk (ERM), and obtain $\hat{h}_n \equiv h_S$

$$\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \hat{L}(h)$$

True Risk=Population Risk

- Let $x \in \mathcal{X}, y \in \mathcal{Y}$, say $\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} \subset \mathbb{R}$, $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- Define a probability measure on $\mathcal{Z}$, denoted by $\mathcal{D}$ in [ML2014] book
- Let $\{z_i = (x_i, y_i)\}_{i=1}^n$ be i.i.d. random variables on a product probability space $\mathcal{Z}^n$.
- Then, for $h \in \mathcal{H}$, we define a true risk/population risk as

$$L(h) \equiv L_{\mathcal{D}}(h) := \mathbb{E}_{x,y}\ell(h(x), y)$$

Probably Approximately Correct (PAC) Learning

The ultimate goal is to find h that minimizes the true risk, i.e.,

$$h \in \arg\min_{h \in \mathcal{H}} L(h)$$

But this is often impossible, leading to the more relaxed definition

Definition (PAC Learnability) A hypothesis class $\mathcal{H}$ is (agnostic) PAC learnable with respect to a set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$, if there exists a learning algorithm, which returns $\hat{h} \equiv \hat{h}(S)$ for a sample S of size $|S| = m$ with the following property: for every $0 < \epsilon, \delta < 1$ there exists $m(\epsilon, \delta)$, such that for all $m \geq m(\epsilon, \delta)$,

$$\mathbb{P}\left[\left\{S \in \mathcal{Z}^m \ : \ L(\hat{h}(S)) \leq \min_{h \in \mathcal{H}} L(h) + \epsilon\right\}\right] \ \geq \ 1 - \delta$$

- PAC learning was introduced by Leslie Valiant (1984), who won for it the Turing Award in 2010.

Learning Infinite Hypothesis Classes

For most hypothesis classes $|\mathcal{H}| = \infty$: What can be done here?

Common measures of complexity of hypothesis classes

- Vapnik-Chervonenkis (VC) dimension
  Chapter 6 in Shai's [ML2014] book
- Rademacher Complexity
  Chapter 26 in [ML2014] book; this was recently used in
  - A Priori Estimates For Two-layer Neural Networks, by Weinan et al., Feb 2020.
  - Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks, by Arora et al., Jan 2019.
- PAC - Bayes: Chapter 31 in [ML20014]
  used recently in several NN papers, e.g.: see Neyshabur et al.
- Compression Bounds: Chapter 30 in [ML2014]

Rademacher Complexity: Motivation

- To simplify the notation, let $f(z) \equiv f(h, z) = \ell(h, z)$ and let $\mathcal{F}$ be the set of these functions
- Then, for $f \in \mathcal{F}$, the population/true risk end empirical risk are equal to

$$L(f) = \mathbb{E}_z[f(z)], \qquad \hat{L}_S(f) = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$$

- Recall ϵ-representative sample: A training set S is called ϵ-representative if

$$\forall h \in \mathcal{H}, \qquad |\hat{L}_S(h) - L(h)| \leq \epsilon \ .$$

- Which motivates the following definition

Definition *Representativeness of S*: is the largest gap between the true error and empirical error

$$\mathrm{Rep}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} |\hat{L}_S(f) - L(f)|$$

Rademacher Complexity: Motivation

- But we don't know the true error $L(f)$
- Replace $L(f)$ by another empirical error
- Partition the training data S into two disjoint sets: S_1, S_2 and estimate the representativeness of S by

$$\sup_{f \in \mathcal{F}} (\hat{L}_{S_1}(f) - \hat{L}_{S_2}(f))$$

- Let $\sigma_i = 1$ if $z_i \in S_1$ and $\sigma_i = -1$, otherwise. Then

$$\sup_{f \in \mathcal{F}} (\hat{L}_{S_1}(f) - \hat{L}_{S_2}(f)) = \frac{1}{m} \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i)$$

  which motivates the following definition

Rademacher Complexity

Definition Let σ_i be i.i.d. with $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = 1/2$. Then, the *Rademacher Complexity* of $\mathcal{F}$ with respect to sample S is defined as

$$R(\mathcal{F}, S) := \frac{1}{m}\mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i)$$

Lemma 26.2 The expected value of the representativeness of S is bounded by

$$\mathbb{E}_S[\mathrm{Rep}(\mathcal{F}, S)] = \mathbb{E}_S\left[\sup_{f\in\mathcal{F}} |\hat{L}_S(f) - L(f)|\right] \leq 2\mathbb{E}_S[R(\mathcal{F}, S)]$$

McDiarmid's Inequality

Lemma

McDiarmid's Inequality Consider independent random random variables $X_1, X_2, \ldots, X_n \in \mathcal{X}$ *and a function* $f : \mathcal{X}^n \to \mathbb{R}$*. If for all* $i \in \{1, \ldots, n\}$ *and all* $x_1, \ldots, x_n, x_n' \in \mathcal{X}$*, the function* f *satisfies*

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1} \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1} \ldots, x_n)| \leq c,$$

then

$$\mathbb{P}[|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| > t] \leq 2 \exp\left(-\frac{2t^2}{nc^2}\right)$$

or, equivalently, with probability at least $1 - \delta$*,*

$$|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \leq c\sqrt{\frac{2}{n} \log\left(\frac{2}{\delta}\right)}.$$

Generalization Bounds

Theorem 26.6 (Shais' book) Assume that for all z and $h \in \mathcal{H}$, we have $|\ell(h,z)| \leq c$, and recall $S = \{z_1, \ldots, z_m\}$ is the sample. Then, with probability at least $1-\delta$, for any $h \in \mathcal{H}$ (and in particular for $h = ERM_{\mathcal{H}}(S)$),

$$L(h) - \hat{L}_S(h) \leq 2\mathbb{E}_S[R(\mathcal{F}, S)] + c\sqrt{\frac{2}{m}\log\left(\frac{2}{\delta}\right)}$$

Proof: First, note that random variable

$$\mathsf{Rep}(\mathcal{F}, S) = \sup_{h\in\mathcal{H}}(L(h) - \hat{L}_S(h))$$

satisfies the bounded difference condition of McDiarmid's lemma with a constant $2c/m$.

Using first the McDiarmid's lemma, and then Lemma 26.2, we obtain that with probability at least $1-\delta$,

$$\mathsf{Rep}(\mathcal{F}, S) \leq \mathbb{E}[\mathsf{Rep}(\mathcal{F}, S)] + c\sqrt{\frac{2}{m}\log\left(\frac{2}{\delta}\right)} \leq 2\mathbb{E}_S[R(\mathcal{F}, S)] + c\sqrt{\frac{2}{m}\log\left(\frac{2}{\delta}\right)}$$

Rademacher Calculus

Hence, to make the preceding theorem useful, we must find the ways to estimate

$$\mathbb{E}_S[R(\mathcal{F}, S)] = ?$$

With a small abuse of notation, from any set $A \in R^m$ let us define

$$R(A) := \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{a}\in A}\sum_{i=1}^{m} a_i\sigma_i\right]$$

This lemma is immediate for the definition. (Why?)

Lemma 26.6 For any $A \subset \mathbb{R}^m$, $c \in \mathbb{R}$, and $\boldsymbol{a}_0 \in \mathbb{R}^m$, we have

$$R(\{c\boldsymbol{a} + \boldsymbol{a}_0 : \boldsymbol{a} \in A\}) \leq |c|R(A).$$

Useful Lemmas

Lemma 26.9 (Contraction Lemma) For each $i \in \{1, \ldots, m\}$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be ρ-Lipschitz, i.e., for any $\alpha, \beta \in \mathbb{R}$, we have $\phi_i(\alpha) - \phi_i(\beta) \leq \rho|\alpha - \beta|$. Furthermore, for any $\in \mathbb{R}$, let $\boldsymbol{\phi}(\boldsymbol{a}) = (\phi_1(a_1), \ldots, \phi_m(a_m))$ and $\boldsymbol{\phi} \circ A = \{\boldsymbol{\phi}(\boldsymbol{a}) : \boldsymbol{a} \in A\}$. Then,

$$R(\boldsymbol{\phi} \circ A) \leq \rho R(A).$$

Rademacher complexity of a finite set $A = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}, \boldsymbol{a}_i \in \mathbb{R}^m$, grows logarithmically with the size of the set $|A| = n$.

Lemma 26.8 (Massart Lemma) Let $\bar{\boldsymbol{a}} = (1/n) \sum_{i=1}^{n} \boldsymbol{a}_i$. Then,

$$R(A) \leq \max_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \bar{\boldsymbol{a}}\| \frac{\sqrt{2 \log(|A|)}}{m}.$$

Rademacher Complexity of Linear Classes

Linear class bounded by L_2 norm. Let

$$\mathcal{H}_2 = \{\boldsymbol{x} \to \langle \boldsymbol{w}, \boldsymbol{x} \rangle : \|\boldsymbol{w}\|_2 \leq 1\}$$

Lemma 26.10 Let $S = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ be vectors in $\boldsymbol{x}_i \in \mathbb{R}^n$. Define $\mathcal{H}_2 \circ S = \{\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle : \|\boldsymbol{w}\|_2 \leq 1, 1 \leq i \leq m\}$. Then,

$$R(\mathcal{H}_2 \circ S) \leq \frac{\max_i \|\boldsymbol{x}_i\|_2}{\sqrt{m}}.$$

Remark: Note that the bound does not depend on the dimension of $\boldsymbol{x}$.
Proof: Using Cauchy-Schwartz inequality $|\boldsymbol{x} \cdot \boldsymbol{y}| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2$,

$$\begin{aligned} mR(\mathcal{H}_2 \circ S) &= \mathbb{E}_\sigma \left[\sup_{\boldsymbol{a} \in \mathcal{H}_2 \circ S} \sum_{i=1}^m \sigma_i a_i \right] = \mathbb{E}_\sigma \left[\sup_{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \leq 1} \sum_{i=1}^m \sigma_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \right] \\ &= \mathbb{E}_\sigma \left[\sup_{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \leq 1} \sum_{i=1}^m \langle \boldsymbol{w}, \sigma_i \boldsymbol{x}_i \rangle \right] \leq \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^m \sigma_i \boldsymbol{x}_i \right\|_2 \right] \end{aligned}$$

Rademacher Complexity of Linear Classes

Proof: Next, using Jensen's inequality and concavity of $\sqrt{x}$

$$\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right\|_2\right] = \mathbb{E}_\sigma\left[\left(\left\|\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right\|_2^2\right)^{1/2}\right] \leq \left(\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right\|_2^2\right]\right)^{1/2}$$

Next

$$\begin{aligned}
\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\right\|_2^2\right] &= \mathbb{E}_\sigma\left[\sum_{i,j} \sigma_i \sigma_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle\right] \\
&= \sum_{i \neq j} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \mathbb{E}[\sigma_i \sigma_j] + \sum_{i=1}^m \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle \mathbb{E}[\sigma_i^2] \\
&= \sum_{i=1}^m \|\boldsymbol{x}_i\|_2^2 \leq m \max_i \|\boldsymbol{x}_i\|_2^2.
\end{aligned}$$

Rademacher Complexity of Linear Classes

Next, using Massart's lemma, we can derive a similar result for L_1 norm. Let

$$\mathcal{H}_1 = \{\boldsymbol{x} \to \langle \boldsymbol{w}, \boldsymbol{x} \rangle : \|\boldsymbol{w}\|_1 \leq 1\}$$

Lemma 26.11 Let $S = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ be vectors in $\boldsymbol{x}_i \in \mathbb{R}^n$. Then,

$$R(\mathcal{H}_1 \circ S) \leq \max_i \|\boldsymbol{x}_i\|_\infty \sqrt{\frac{2\log(2n)}{m}}$$

Proof: Using $\langle \boldsymbol{w}, \boldsymbol{v} \rangle \leq \|\boldsymbol{w}\|_1 \|\boldsymbol{v}\|_\infty$,

$$\begin{aligned} mR(\mathcal{H}_1 \circ S) &= \mathbb{E}_\sigma \left[\sup_{\boldsymbol{a} \in \mathcal{H}_1 \circ S} \sum_{i=1}^m \sigma_i a_i \right] = \mathbb{E}_\sigma \left[\sup_{\boldsymbol{w} : \|\boldsymbol{w}\|_1 \leq 1} \sum_{i=1}^m \sigma_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \right] \\ &= \mathbb{E}_\sigma \left[\sup_{\boldsymbol{w} : \|\boldsymbol{w}\|_1 \leq 1} \sum_{i=1}^m \langle \boldsymbol{w}, \sigma_i \boldsymbol{x}_i \rangle \right] \leq \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^m \sigma_i \boldsymbol{x}_i \right\|_\infty \right] \end{aligned}$$

Rademacher Complexity of Linear Classes

Proof: Next, for $1 \leq j \leq n$, let $\boldsymbol{v}_j = (x_{1,j}, \ldots, x_{m,j})$ be the jth coordinate of all $\boldsymbol{x}$ vectors. Note that

$$\|\boldsymbol{v}_j\|_2 \leq \sqrt{m} \max_i \|\boldsymbol{x}\|_\infty \quad \text{(why?)}$$

Now, let $V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n, -\boldsymbol{v}_1, \ldots, -\boldsymbol{v}_n\}$. Then, observe

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^{m} \sigma_i \boldsymbol{x}_i \right\|_\infty \right] = mR(V) \leq m \max_i \|\boldsymbol{x}_i\|_\infty \sqrt{\frac{2\log(2n)}{m}},$$

where in the last inequality we used Massart's lemma.

A Priori Estimates For Two-layer NNs

Paper:

- A Priori Estimates For Two-layer Neural Networks, Weinan et al., 2019.

The consider two ways of bounding the error (in some norm):

- f^*: true function that we want to estimate
- $\hat{f}_n$: numerical estimate based on a sample of size n
- A priori error estimate
$$\|\hat{f}_n - f^*\| = O(\|f^*\|)$$
- A posteriori error estimate
$$\|\hat{f}_n - f^*\| = O(\|\hat{f}_n\|)$$
- In this context, most of the recent work on generalization error of NNs can be viewed as "a posteriori".
Values of $\|\hat{f}_n\|$ are often huge, yielding often vacuous bounds.

Paper Notation

- Training set: $S = \{(x_i, y_i)\}_{i=1}^n$; i.i.d.samples from a distribution $\rho_{x,y}$
- True (target) function: $f^*(x) = \mathbb{E}[y|x]$ where $y = f(x) + \xi$ with ξ being the noise.
- $f^*(x) : [-1, 1]^d \to [0, 1]$
- One hidden layer neural network

$$f(x;\theta) = \sum_{k=1}^{m} a_k \sigma(w_k^\top x)$$

  where $w_k \in \mathbb{R}^d, a_k \in \mathbb{R}$ and $\theta = \{(a_k, w_k)\}_{k=1}^m$
- $\sigma(x) : \mathbb{R} \to \mathbb{R}$: activation function
  $\sigma(x)$ is 1-Lipschitz
  $\sigma(x)$ scale-free: $\sigma(\alpha x) = \alpha\sigma(x), \alpha \geq 0, x \in \mathbb{R}$
  e.g., ReLU or Leaky ReLU

Training

- Loss function: $\ell(y, y') = (y - y')^2/2$
- Ultimate goal: minimize the population (true) risk

$$L(\theta) = \mathbb{E}_{x,y}[\ell(f(x;\theta), y)]$$

- In practice: minimize the empirical risk

$$\hat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;\theta), y_i)$$

- Regularized empirical risk:

$$J_\lambda(\theta) := \hat{L}_n(\theta) + \lambda(\|\theta\|_{\mathcal{P}} + 1)$$

where $\|\theta\|_{\mathcal{P}} := \sum_{k=1}^{m} |a_k|\|w_k\|$ is the path norm (Neyshabur et al., 2015)

- Regularized ERM: $$\hat{\theta}_{n,\lambda} = \arg\min_\theta J_\lambda(\theta)$$

where $\lambda > 0$ is the tuning parameter.

Barron Space

Based on the work by Barron (1993).

- Barron function: A function $f : \Omega \to \mathbb{R}$ is called a Barron if it admits the following representation

$$f(x) = \int_{S^d} a(w)\sigma(w^\top x)d\pi(w),$$

  where π is a probability distribution over $S^d = \{w : \|w\|_1 = 1\}$ and $a(\cdot)$ is a scalar function.

- Barron norm: Let f be a Barron function.
  - Denote by Θ_f all possible representations of f

$$\Theta_f = \left\{(a, \pi) : f(x) = \int_{S^d} a(w)\sigma(w^\top x)d\pi(w)\right\}$$

  - Barron norm $\gamma_p(f)$:

$$\gamma_p(f) := \inf_{(a,\pi)\in\Theta_f} \left(\int_{S^d} |a(w)|^p d\pi(w)\right)^{1/p}$$

Barron Space

- Barron space
$$\mathcal{B}_p(\Omega) = \{f(x) : \gamma_p(f) < \infty\}$$
- Since π is a probability distribution, by Hölder's inequality
$$\gamma_p(f) \le \gamma_q(f), \qquad \text{if } q \ge p > 0.$$
and thus
$$\mathcal{B}_\infty(\Omega) \subset \cdots \subset \mathcal{B}_2(\Omega) \subset \mathcal{B}_1(\Omega)$$

Theorem 3.1 For any $f \in \mathcal{B}_2(\Omega)$, there exists a 2-layer NN $f(x;\tilde{\theta})$ of width m with $\|\tilde{\theta}\|_{\mathcal{P}} = \sum_{k=1}^m |a_k|\|w_k\| \le 2\gamma_2(f)$, such that
$$\mathbb{E}_x(f(x) - f(x;\tilde{\theta}))^2 \le \frac{3\gamma_2^2}{m}.$$

Intuition: Integral representation ensures a good approximation
$$f(x) \approx \frac{1}{m}\sum_{k=1}^m a(w_k)\sigma(w_k^\top x)$$

Proof of Theorem 3.1

- By assumption, let (a, π) be the best representation of f, such that the Barron-2 norm is

$$\gamma_2^2(f) = \mathbb{E}_\pi[(a(w))^2].$$

- Let $U = \{w_j\}_{j=1}^m$ be i.i.d. random variables with distribution $\pi(\cdot)$ and

$$\hat{f}_U(x) = \frac{1}{m}\sum_{j=1}^m a(w_j)\sigma(x^\top w_j).$$

- Let $L_U = \mathbb{E}_x[(\hat{f}_U(x) - f(x))^2]$ be the population risk, and

$$\begin{aligned}
&\mathbb{E}_U[L_U] = E_x E_U[(\hat{f}_U(x) - f(x))^2] \\
&= \frac{1}{m^2}\mathbb{E}_x \sum_{j,l} \mathbb{E}_{w_j,w_l}[(a(w_j)\sigma(x^\top w_j) - f(x))(a(w_l)\sigma(x^\top w_l) - f(x))] \\
&= \frac{1}{m}\mathbb{E}_x\mathbb{E}_w[(a(w)\sigma(x^\top w) - f(x))^2] \le \frac{\gamma_2^2(f)}{m}
\end{aligned}$$

Proof of Theorem 3.1

- Next, let A_U be the path norm of $\hat{f}_U$ and note

$$\mathbb{E}_U[A_U] = \gamma_1(f) \leq \gamma_2(f)$$

- Then, we define events

$$E_1 = \left\{ L_U < \frac{3\gamma_2^2(f)}{m} \right\} \quad \text{and} \quad E_2 = \{A_U < 2\gamma_1(f)\}$$

  and use the Markov's inequality

$$\mathbb{P}[E_1] = 1 - \mathbb{P}\left[L_U \geq \frac{3\gamma_2^2(f)}{m} \right] \geq 1 - \frac{\mathbb{E}_U[L_U]}{3\gamma_2^2(f)/m} \geq \frac{2}{3}$$

$$\mathbb{P}[E_2] = 1 - \mathbb{P}[A_U \geq 2\gamma_2(f)] \geq 1 - \frac{\mathbb{E}[A_U]}{2\gamma_2(f)} \geq \frac{1}{2}$$

  Therefore, the following completes the proof

$$\mathbb{P}[E_1 E_2] \geq \mathbb{P}[E_1] + \mathbb{P}[E_2] - 1 \geq \frac{2}{3} + \frac{1}{2} - 1 > 0$$

Main Result: Generalization Bound

Noiseless clase: $\xi = 0$; Also, assume $\ln(2d) \geq 2$ and $\hat{\gamma}_2 = \max(1, \gamma_2(f))$.

Theorem 4.1 Assume $f^* \in \mathcal{B}_2(\Omega)$ and $\lambda \geq \lambda_n := 4\sqrt{2\ln(2d)/n}$, where n is the number of samples. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of the training set S,

$$\mathbb{E}_x[(f(x;\tilde{\theta}_{n,\lambda}) - f^*(x))^2] \lesssim \frac{\gamma_2^2(f^*)}{m} + \lambda\hat{\gamma}_2(f^*) \quad (1)$$

$$+ \frac{1}{\sqrt{n}}\left(\hat{\gamma}_2(f^*) + \sqrt{\ln(n/\delta)}\right) \quad (2)$$

Remark Note that if we choose $\lambda = 4\sqrt{2\ln(2d)/n}$ and $m \geq \sqrt{n}$, then

$$\mathbb{E}_x[(f(x;\tilde{\theta}_{n,\lambda}) - f^*(x))^2] = O\left(\frac{1}{\sqrt{n}}\right)$$

Comparison With Kernel Methods

- For fixed π, the following is a Reproducing Kernel Hilbert Space (RKHS)

$$\mathcal{H}_\Omega = \left\{ f = \int_{S^d} a(w)\sigma(x^\top w)d\pi(w) : \|f\|_{\mathcal{H}_\pi} < \infty \right\}$$

  where

$$\|f\|^2_{\mathcal{H}_\pi} := \mathbb{E}_\pi[|a(w)|^2]$$

  The corresponding kernel (see Rahimi&Rechet, 2008) is defined by

$$k_\pi(x, x') = \mathbb{E}_\pi[\sigma(x^\top w)\sigma(x'^\top w)].$$

- Note that Barron's space is much reacher

$$\mathcal{B}_2(\omega) = \cup_\pi \mathcal{H}_\pi(\Omega)$$

Comparison With Kernel Methods

More formally

- Consider $f^* \in \mathcal{B}_2(\omega)$ and let a^*, π^* be its best representation, i.e.

$$\gamma_2^2(f^*) = \mathbb{E}_{\pi^*}[|a^*(w)|^2]$$

- For fixed π_0, if π^* is absolutely continuous withe respect to π_0

$$\begin{aligned} f^* &= \int_{S^d} a^*(w)\sigma(x^\top w) d\pi^*(w) \\ &= \int_{S^d} a^*(w) \frac{d\pi^*}{d\pi_0} \sigma(x^\top w) d\pi_0(w) \end{aligned}$$

and

$$\|f\|^2_{\mathcal{H}_{\pi_0}} = \mathbb{E}_{\pi_0}[|a^*(w)) \frac{d\pi^*}{d\pi_0}|^2] \geq \gamma_2(f^*)$$

- The best generalization bound using kernel k_{π_0} is (Caponnetto&De Vito, 2007) of the order

$$\frac{\|f\|_{\mathcal{H}_{\pi_0}}}{\sqrt{n}} \geq \frac{\gamma_2(f^*)}{\sqrt{n}}$$

which is comparable to Theorem 4.1

A Posterior Bound

The proof of Theorem 4.1 uses the following a posterior bound.

Theorem 5.2 (A posterior generalization bound) Assume that the loss function $\ell(\cdot, y)$ is A-Lipschitz continuous and bounded by B. Then, for any δ, with probability at least $1-\delta$, over the choice of the training set S, we have, for any 2-layer NN

$$|L(\theta) - \hat{L}_n(\theta)| \leq 4A\sqrt{\frac{2\ln(2d)}{n}}(\|\theta\|_{\mathcal{P}} + 1) + B\sqrt{\frac{2\ln(2c(\|\theta\|_{\mathcal{P}} + 1)^2/\delta)}{n}},$$

where $c = \sum_{k=1}^{\infty} 1/k^2$

Note that the generalization gap is roughly bounded by $\|\theta\|_{\mathcal{P}}/\sqrt{n}$

The **proof** is based on Theorem 26.5 in Shais' book, which we covered, and the next lemma.

Rademacher Complexity of 2-Layer NN

Lemma B.3 Let $\mathcal{F}_C = \{f_m(x;\theta) | \|\theta\|_{\mathcal{P}} \leq C\}$ be the set of 2-L NN with path norm bounded by C. Then

$$R_n(\mathcal{F}_C) \leq 2C\sqrt{\frac{2\ln(2d)}{n}}$$

Proof relies on Lemmas 26.9 & 26.11 from Shais' book, which we covered before.

Let ξ_i be Rademacher i.i.d. random variables with $\mathbb{P}[\xi_i = \pm 1] = 1/2$ and $\xi = (\xi_1, \ldots, \xi_n)$

Rademacher Complexity of 2-Layer NN

Proof

$$
\begin{aligned}
nR_n(\mathcal{F}_C) &= \mathbb{E}\left[\sup_{\|\theta\|_{\mathcal{P}}\le C} \sum_{i=1}^{n} \xi_i \sum_{k=1}^{m} a_k \sigma(x_i^\top w_k)\right] \\
&\le \mathbb{E}_\xi\left[\sup_{\|\theta\|_{\mathcal{P}}\le C, \|u_k\|_1=1} \sum_{i=1}^{n} \xi_i \sum_{k=1}^{m} a_k \|w_k\|_1 \sigma(x_i^\top u_k)\right] \\
&= \mathbb{E}_\xi\left[\sup_{\|\theta\|_{\mathcal{P}}\le C, \|u_k\|_1=1} \sum_{k=1}^{m} a_k \|w_k\|_1 \sum_{i=1}^{n} \xi_i \sigma(x_i^\top u_k)\right] \\
&\le \mathbb{E}_\xi\left[\sup_{\|\theta\|_{\mathcal{P}}\le C} \sum_{k=1}^{m} a_k \|w_k\|_1 \sup_{\|u\|_1=1} \left|\sum_{i=1}^{n} \xi_i \sigma(x_i^\top u)\right|\right] \\
&\le C\mathbb{E}_\xi\left[\sup_{\|u\|_1\le 1} \left|\sum_{i=1}^{n} \xi_i \sigma(x_i^\top u)\right|\right]
\end{aligned}
$$

Finally, use 1-Lipschitz continuity of σ and apply Lemmas 26.9 & 26.11 from the [ML] book.

Proof of Theorem 5.2

- Decompose the hypothesis class, $\mathcal{F}$, into $\mathcal{F} = \cup_{l=1}^{\infty} \mathcal{F}_l$, where

$$\mathcal{F}_l = \{f_m(x;\theta) | \|\theta\|_{\mathcal{P}} \leq l\}$$

- Set $\delta_l = \delta/(cl^2)$, where $c = \sum_{l=1}^{\infty} 1/l^2$
- Apply Theorem 26.5 from the [ML] book and the preceding Lemma B.3 to each hypothesis class $\mathcal{F}_l$
- Combining all the bounds, we obtain that with probability at least $1-\delta$, $\delta = \sum \delta_l$, the following inequality holds for all l,

$$\sup_{\|\theta_{\mathcal{P}}\| \leq l} |L(\theta) - \hat{L}_n(\theta)| \leq 4Al\sqrt{\frac{2\ln(2d)}{n}} + B\sqrt{\frac{2\ln(2/\delta_l)}{n}}$$

- Pick $l_0 = \min\{l : \|\theta_{\mathcal{P}}\| \leq l\}$ and then use $l_0 \leq \|\theta_{\mathcal{P}}\| + 1$

Proof of Main Theorem 4.1

- The main idea is to bound $\|\hat{\theta}_n\|_{\mathcal{P}}$ by a well behaved path norm $\|\tilde{\theta}\|_{\mathcal{P}}$ from Theorem 3.1 since $\|\tilde{\theta}\|_{\mathcal{P}} \leq 2\gamma_2(f^*)$.
- From Theorem 5.2 we have with probability at least $1-\delta$

$$L(\hat{\theta}_{n,\lambda}) \leq \hat{L}(\hat{\theta}_{n,\lambda}) + \lambda_n(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} + 1) + 3\sqrt{\frac{\ln(2c(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} + 1)^2/\delta)}{n}}$$

$$\leq J_\lambda(\hat{\theta}_{n,\lambda}) + 3\sqrt{\frac{\ln(2c(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} + 1)^2/\delta)}{n}}, \tag{3}$$

  where the last inequality is due to the choice of $\lambda \geq \lambda_n = 4\sqrt{2\ln(2d)/n}$
- The first term can be bounded as

$$J_\lambda(\hat{\theta}_{n,\lambda}) \leq J_\lambda(\tilde{\theta}),$$

  which follows from the definition of $\hat{\theta}_{n,\lambda}$, where $\tilde{\theta}$ is the same as in Theorem 3.1

Proof of Main Theorem 4.1

- Then, recalling $\lambda_n = 4\sqrt{2\ln(2d)/n}$ and the claim of Theorem 3.1,

$$\begin{aligned} J(\tilde{\theta}) &= \hat{L}(\tilde{\theta}) + \lambda(\|\tilde{\theta}\|_{\mathcal{P}} + 1) \\ &\leq L(\tilde{\theta}) + (\lambda_n + \lambda)(\|\tilde{\theta}\|_{\mathcal{P}} + 1) \\ &\leq L(\tilde{\theta}) + 6\lambda\gamma_2(f^*) + 2\sqrt{\frac{2\ln(2c(1+2\gamma_2(f^*))^2/\delta}{n}} \\ &\leq L(\tilde{\theta}) + 8\lambda\gamma_2(f^*) + 2\sqrt{\frac{\ln(2c/\delta)}{n}} \end{aligned}$$

  where in second to the last inequality we used $\|\tilde{\theta}\|_{\mathcal{P}} \leq 2\gamma_2(f^*)$, and in the last $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \ln(1+a) \leq a, a \geq 0, b \geq 0$. (The preceding inequality is stated as Proposition 5.1.)

- Moreover,

$$\sqrt{n}\sqrt{\frac{\ln(2c(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} + 1)^2/\delta)}{n}} \leq \sqrt{\ln(2nc/\delta)} + \sqrt{2\frac{\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}}{\sqrt{n}}}$$

- Next, $\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}$ is bounded by Proposition 5.2, and putting all the bounds together and simplifying, yields the result.

Proof of Main Theorem 4.1

- Next, $\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}$ can be bounded using

$$\lambda(\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}+1) \leq J_\lambda(\hat{\theta}_{n,\lambda}) \leq J_\lambda(\tilde{\theta}) \leq L(\tilde{\theta})+8\lambda\gamma_2(f^*)+2\sqrt{\frac{\ln(2c/\delta)}{n}},$$

  which we showed earlier, implying

$$\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}} \leq \frac{L(\tilde{\theta})}{\lambda} + 8\hat{\gamma}_2(f^*) + \frac{1}{2}\sqrt{\ln(2c/\delta)}.$$

- Finally, combining the preceding bounds in (4), yields

$$L(\hat{\theta}_{n,\lambda}) \lesssim L(\tilde{\theta})+8\lambda\hat{\gamma}_2(f^*)+\frac{3}{\sqrt{n}}\left(\sqrt{\frac{L(\tilde{\theta})}{n^{1/2}\lambda}} + \hat{\gamma}_2(f^*) + \sqrt{\ln(n/\delta)}\right),$$

  which, in combination with $L(\tilde{\theta}) \leq 3\gamma_2^2(f^*)/m$ from Theorem 3.1, yields the proof.

Reading On Overparametrization

Lazy regime: weights don't move much. The following paper pinpoints the key ingredient responsible for wights remaining close to the initial values during training

- A Note on Lazy Training in Supervised Differentiable Programming, by L. Chizat and F. Bach, Dec 2018.

  See also an updated version:
  On Lazy Training Differentiable Programming, by L. Chizat and F. Bach, NIPS 2019.
- Chapters 13&14 in the recent monograph Deep Learning Theory Lecture Notes, by Telgarsky, Feb 2021.

Mean Filed regime: weight move a lot.

- On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport, by L. Chizat and F. Bach, NIPS 2018.
- A Mean Field View of the Landscape of Two-Layers Neural Networks, Mei, Song, Andrea Montanari, and Phan-Minh Nguyen, 2018.

See the references in the preceding papers, and the citations on Google Scholar

Reading On Generalization

Generalization bounds:

- PAC Learning and Generalization Theory - [ML2014] book:
  PAC learning: Chapters 2-4; Rademacher Complexity: Chapter 26
  (In particular, Theorem 26.5 and Lemmas 26.9 & 26.11)
  In general, for PAC learning theory see: Chapters: 2-6, 26-31
- Generalization bounds for NNs
  - A Priori Estimates For Two-layer Neural Networks, by Weinan et al., Feb 2020..
  - Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks, by Arora et al., Jan 2019.
  - See the references in the preceding papers, and the citations on Google Scholar

Have Fun!