A Function Space Theory and Generalization Error Estimates for Neural Network Models

Chao Ma

Online Summer School of Deep Learning Theory

Shanghai Jiao Tong University

July 17, 2020

Table of Contents

What problem can be solved?

- Neural Networks can solve many problems: CV, NLP, RL, ...
- But cannot solve any problem

$$A\boldsymbol{x} = \boldsymbol{y}$$

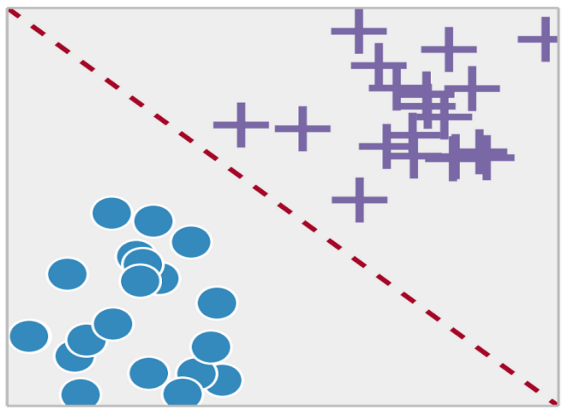
We have to characterize the set of problems (target functions) that neural network models can learn.

Three steps for supervised learning

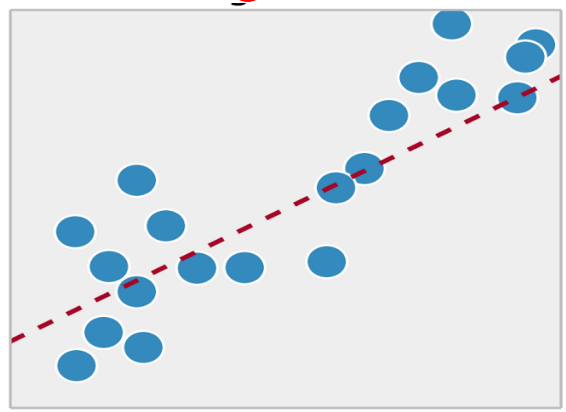
Supervised learning problems:

- Given data $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, $\boldsymbol{x}_i \sim \mu^*$, $y_i = f^*(\boldsymbol{x}_i) + \epsilon_i$
- Learn the relation between $\boldsymbol{x}$ and y (the target function f^*) from $\mathcal{S}$

Classification



Regression



Three steps for supervised learning problems:

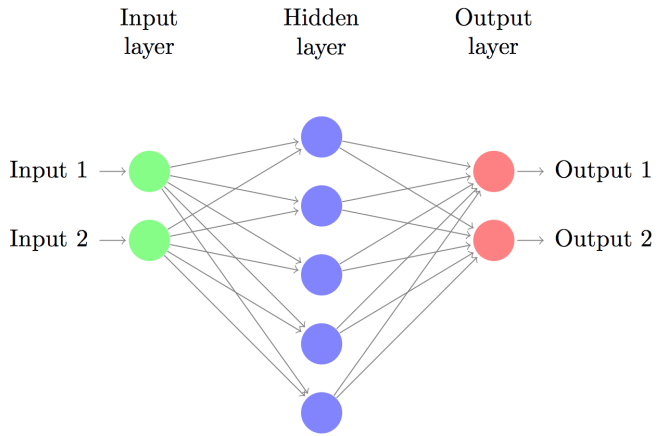
1. Construct a hypothesis space $\mathcal{H}$ with a parameterized model
2. Design a loss function
3. Take an optimization algorithm to minimize the loss function

Step 1: hypothesis space by parameterized model

- Construct a parameterized model $f_\theta(\boldsymbol{x})$
- The hypothesis space $\mathcal{H} = \{f_\theta(\cdot) : \ \theta \in \Omega\}$

Linear model:

$$f_\theta(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m} a_k \phi_k(\boldsymbol{x})$$



Two-layer neural networks:

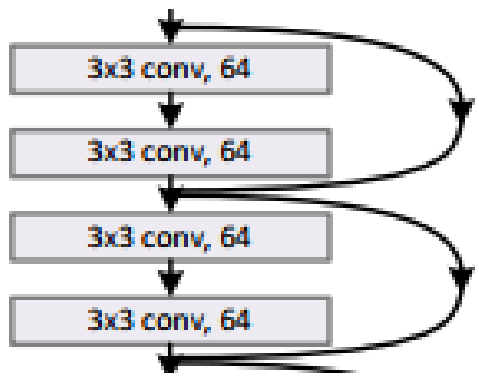
$$f_\theta(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m} a_k \sigma(\boldsymbol{b}_k^T \boldsymbol{x} + c_k)$$

Residual networks:

$$\boldsymbol{z}_0(\boldsymbol{x}) = \boldsymbol{V}\boldsymbol{x}$$

$$\boldsymbol{z}_{l+1}(\boldsymbol{x}) = \boldsymbol{z}_l(\boldsymbol{x}) + \frac{1}{L}\boldsymbol{U}_l \sigma \circ (\boldsymbol{W}_l z_{l,L}(\boldsymbol{x}))$$

$$f_\theta(\boldsymbol{x}) = \boldsymbol{\alpha}^T \boldsymbol{z}_L(\boldsymbol{x})$$



Step 2: loss function

- The loss function measures the closeness between f_θ and f^*
- It can only depend on training data $\mathcal{S}$

$$\text{Loss}(\theta) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} l(f_\theta(\boldsymbol{x}_i), y_i)}_{\text{empirical loss}} + \underbrace{reg(\theta)}_{\text{regularization}}$$

Choice of l:

square $(f_\theta(\boldsymbol{x}) - y)^2$, cross entropy $-log(f_\theta^{(c)}(\boldsymbol{x}))$, etc.

Step 3: finding the solution

Solve the minimizing problem

$$\min_{\theta \in \Omega} \mathrm{Loss}(\theta)$$

Algorithms:

- Gradient Descent (GD): $\theta_{t+1} = \theta_t - \eta \nabla \mathrm{Loss}(\theta_t)$
- (mini-batch) Stochastic Gradient Descent (SGD):

$$\theta_{t+1} = \theta_t - \frac{\eta}{B} \sum_{i=1}^{B} \nabla l(f_\theta(\boldsymbol{x}_{b_i}), y_{b_i})$$

- Momentum GD, Nesterov accelerated GD, Adaptive algorithms, etc.

Theoretical issues for supervised learning

Three theoretical issues corresponding to the three steps:

- **Constructing hypothesis space** ..... **Approximation**
  Is the hypothesis space large enough to approximate the target function with small accuracy?
- **Formulating the loss function** ..... **Generalization**
  How to estimate the performance of the model on unseen data (population risk), while only training data are accessible?
- **Finding the solution** ..... **Optimization**
  Does the optimization algorithm converge? Does the solution found by the algorithm generalize well?

We focus on two issues: Approximation and Estimation

Analogy in Numerical Analysis

In Numerical Analysis:

1. Approximation theories:

$$\inf_{f_m} \|f_m - f^*\| \leq \frac{C|f^*|_{W^r}}{m^r}.$$

2. Estimation: (FEM)

$$\|f_m - f^*\|_{L^p} \leq Cm^{-\alpha}\|f^*\|_{W^{s,p}}$$

Similar theories for neural network models?

Function spaces

- Banach Space, Hilbert Space
- Continuous functions, Lipschitz functions, L^p space, Sobolev Space, Besov Space, ...
- Reproducing Kernel Hilbert Space (RKHS)

The curse of dimensionality

Universal approximation theorem: let $f_m(\cdot;\theta)$ be two-layer neural networks with m neurons and parameters θ, then for any continuous function f^*,

$$\lim_{m\to\infty} \inf_{\theta} \|f_m(\cdot;\theta) - f^*(\cdot)\| = 0$$

The curse of dimensionality

Universal approximation theorem: let $f_m(\cdot;\theta)$ be two-layer neural networks with m neurons and parameters θ, then for any continuous function f^*,

$$\lim_{m\to\infty} \inf_{\theta} \|f_m(\cdot;\theta) - f^*(\cdot)\| = 0$$

Is the space of continuous function enough?

The curse of dimensionality

Universal approximation theorem: let $f_m(\cdot;\theta)$ be two-layer neural networks with m neurons and parameters θ, then for any continuous function f^*,

$$\lim_{m\to\infty} \inf_{\theta} \|f_m(\cdot;\theta) - f^*(\cdot)\| = 0$$

Is the space of continuous function enough?

NO! The Curse of Dimensionality

$$\|f_m - f^*\|_{L^p} \leq Cm^{-\alpha/d}\|f^*\|_{W^{s,p}}$$

we may need $m > \varepsilon^{-d}$ to achieve ε approximation error.

The curse of dimensionality is intrinsic for high dimensional spaces

The curse of dimensionality is intrinsic for high dimensional spaces

We can only avoid it by considering a smaller set of problems

Overcome CoD: Barron's theorem

Theorem (Barron '93)

For function $f(x) : [0,1]^d \to \mathbb{R}$, *let* $\hat{f}(\omega)$ *be the Fourier transform of any extension of* f *to* $\mathbb{R}^d$. *Then, if* $\gamma(f) := \int \|\omega\|_1^2 |\hat{f}(\omega)| d\omega < \infty$, *for any* $m > 0$ *there exists a two-layer neural network* $f_m(x) = \sum_{k=1}^m a_k \sigma(b_k^T x + c_k)$, *with* $\sigma(\cdot)$ *being the RELU activation function, satisfying*

$$\|f_m(x) - f(x)\|^2 \leq \frac{3\gamma(f)^2}{m},$$

and $\sum_{k=1}^m |a_k|(\|b_k\| + |c_k|) \leq 2\gamma(f)$.

Overcome CoD: RKHS

- Features: $\phi(\boldsymbol{x}, \boldsymbol{v})$
- Kernel:
$$K(\boldsymbol{x}, \boldsymbol{x}') = \int \phi(\boldsymbol{x}, \boldsymbol{v})\phi(\boldsymbol{x}', \boldsymbol{v})\mu(d\boldsymbol{v})$$
- Reproducing Kernel Hilbert Space (RKHS):
$$\mathcal{H} := \left\{ \int a(\boldsymbol{v})\phi(\boldsymbol{x}, \boldsymbol{v})\mu(d\boldsymbol{v}) \right\}$$
Norm: $\|f\|_{\mathcal{H}} := \int |a(\boldsymbol{v})|^2 \mu(d\boldsymbol{v})$
- Another definition:
$$\mathcal{H} := \left\{ \sum_{k=1}^{m} a_k K(\boldsymbol{x}, \boldsymbol{x}_k) \right\}$$

Overcome CoD: RKHS

The random feature model:

$$\hat{f}(\boldsymbol{x}, \boldsymbol{a}) = \sum_{k=1}^{m} a_k \phi(\boldsymbol{x}, \boldsymbol{v}_i), \qquad \boldsymbol{v}_i \sim \mu$$

- Two-layer neural networks with the first layer fixed
- Both approximation and estimation without CoD

$$\inf_{\boldsymbol{a}} \|\hat{f}(\boldsymbol{x}, \boldsymbol{a}) - f(\boldsymbol{x})\|_{L^2} \lesssim \frac{\|f\|_{\mathcal{H}}}{\sqrt{m}}, \quad w.h.p$$

RKHS and L^2

Decomposition of K:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=0}^{\infty} \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{x}')$$

If $\{\psi_i\}$ is a complete bases, then

$$\|f\|_{L^2}^2 = \sum_{i=0}^{\infty} \langle f, \psi_i \rangle^2,$$

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=0}^{\infty} \lambda_i^{-1} \langle f, \psi_i \rangle^2$$

Table of Contents

In this Section

- Barron space for two-layer neural networks, and its properties
- Approximation results for functions in the Barron space, by two-layer neural networks
- A priori estimates of generalization errors for functions in the Barron space, for two-layer neural networks and residual networks.

The Barron space

Two-layer neural networks:

$$f_m(\boldsymbol{x};\theta) = \frac{1}{m}\sum_{k=1}^{m} a_k \sigma(\boldsymbol{b}_k^T \boldsymbol{x} + c_k), \quad \boldsymbol{x} \in [0,1]^d$$

Takes similar for as an empirical approximation (Monte-Carlo) of an expectation.
Consider the function $f : [0,1]^d \mapsto \mathbb{R}$ of the following integral form

$$f(\boldsymbol{x}) = \int_{\Omega} a\sigma(\boldsymbol{b}^T \boldsymbol{x} + c)\rho(da, d\boldsymbol{b}, dc), \quad \boldsymbol{x} \in [0,1]^d$$

$\Omega = \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$, ρ is a probability distribution on Ω.

The Barron Space

$$f(\boldsymbol{x}) = \int_{\Omega} a\sigma(\boldsymbol{b}^T\boldsymbol{x} + c)\rho(da, d\boldsymbol{b}, dc), \quad \boldsymbol{x} \in [0,1]^d$$

Take the "moment" of ρ as norm of functions:

$$\|f\|_{\mathcal{B}_2} = \inf_{\rho \in P_f} \sqrt{\mathbb{E}_\rho[a^2(\|\boldsymbol{b}\|_1 + |c|)^2]},$$

where $P_f := \{\rho : f(\boldsymbol{x}) = \mathbb{E}_\rho[a\sigma(\boldsymbol{b}^T\boldsymbol{x} + c)]\}$. Let the Barron-2 space be

$$\mathcal{B}_2 = \{f \in C^0 : \|f\|_{\mathcal{B}_2} < \infty\}$$

Why named "Barron Space"

Early study of approximation without the curse of dimensionality:

Theorem (Barron '93)

For function $f(x) : [0,1]^d \to \mathbb{R}$, let $\hat{f}(\omega)$ be the Fourier transform of any extension of f to $\mathbb{R}^d$. Then, if $\gamma(f) := \int \|\omega\|_1^2 |\hat{f}(\omega)| d\omega < \infty$, for any $m > 0$ there exists a two-layer neural network $f_m(x) = \sum_{k=1}^m a_k \sigma(b_k^T x + c_k)$, with $\sigma(\cdot)$ being the RELU activation function, satisfying

$$\|f_m(x) - f(x)\|^2 \leq \frac{3\gamma(f)^2}{m},$$

and $\sum_{k=1}^m |a_k|(\|b_k\| + |c_k|) \leq 2\gamma(f).$

Considered function class:

$$\{f : \gamma(f) < \infty\}$$

Barron Space contains $\{f : \gamma(f) < \infty\}$

Theorem

Let $f \in C(X)$, the space of continuous functions on X, and assume that f satisfies $\gamma(f) < \infty$. Then f admits an integral representation

$$f(\boldsymbol{x}) = \int_{\Omega} a\sigma(\boldsymbol{b}^T \boldsymbol{x} + c)\rho(da, d\boldsymbol{b}, dc)$$

Moreover,

$$\|f\|_{\mathcal{B}_2} \leq 2\gamma(f) + 2\|\nabla f(0)\|_1 + 2|f(0)|.$$

Proved in Barron 93

Index of Barron Spaces

Consider $f(x) : [0,1]^d \to \mathbb{R}$, let

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho \in P_f} \left(\mathbb{E}_\rho |a|^p (\|b\|_1 + |c|)^p\right)^{1/p}.$$

Define Barron space $\mathcal{B}_p$ as

$$\mathcal{B}_p = \{f : \ \|f\|_{\mathcal{B}_p} < \infty\}.$$

Index of Barron Spaces

Consider $f(x) : [0,1]^d \to \mathbb{R}$, let

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho \in P_f} \left(\mathbb{E}_\rho |a|^p (\|b\|_1 + |c|)^p\right)^{1/p}.$$

Define Barron space $\mathcal{B}_p$ as

$$\mathcal{B}_p = \{f : \ \|f\|_{\mathcal{B}_p} < \infty\}. \qquad \mathcal{B}_\infty \subset \cdots \subset \mathcal{B}_2 \subset \mathcal{B}_1$$

Index of Barron Spaces

Consider $f(x) : [0,1]^d \to \mathbb{R}$, let

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho \in P_f} \left(\mathbb{E}_\rho |a|^p (\|b\|_1 + |c|)^p\right)^{1/p}.$$

Define Barron space $\mathcal{B}_p$ as

$$\mathcal{B}_p = \{f : \ \|f\|_{\mathcal{B}_p} < \infty\}. \qquad \mathcal{B}_\infty \subset \cdots \subset \mathcal{B}_2 \subset \mathcal{B}_1$$

Theorem

For any $f \in \mathcal{B}_1$, we have $f \in \mathcal{B}_\infty$ and

$$\|f\|_{\mathcal{B}_1} = \|f\|_{\mathcal{B}_\infty}.$$

Index of Barron Spaces

Consider $f(x) : [0,1]^d \to \mathbb{R}$, let

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho \in P_f} \left(\mathbb{E}_\rho |a|^p (\|b\|_1 + |c|)^p\right)^{1/p}.$$

Define Barron space $\mathcal{B}_p$ as

$$\mathcal{B}_p = \{f : \ \|f\|_{\mathcal{B}_p} < \infty\}. \qquad \mathcal{B}_\infty \subset \cdots \subset \mathcal{B}_2 \subset \mathcal{B}_1$$

Theorem

For any $f \in \mathcal{B}_1$, we have $f \in \mathcal{B}_\infty$ and

$$\|f\|_{\mathcal{B}_1} = \|f\|_{\mathcal{B}_\infty}.$$

$$\mathcal{B}_1 = \mathcal{B}_\infty = \mathcal{B}$$

Brief proof

For $f \in \mathcal{B}_1$:

$$f(\boldsymbol{x}) = \int_{\Omega} a\sigma(\boldsymbol{b}^T\boldsymbol{x} + c)\rho(da, d\boldsymbol{b}, dc), \quad \forall\, \boldsymbol{x} \in D_0, \tag{1}$$

and

$$\mathbb{E}_{\rho}\left[|a|(\|\boldsymbol{b}\|_1 + |c|)\right] < \infty. \tag{2}$$

Let $\Lambda = \{(\boldsymbol{b}, c) : \ \|\boldsymbol{b}\|_1 + |c| = 1\}$, and consider two measures ρ_+ and ρ_- on Λ defined by

$$\rho_+(A) = \int_{\{(a,\boldsymbol{b},c):\ (\hat{\boldsymbol{b}},\hat{c})\in A, a>0\}} |a|(\|\boldsymbol{b}\|_1 + |c|)\rho(da, d\boldsymbol{b}, dc), \tag{3}$$

$$\rho_-(A) = \int_{\{(a,\boldsymbol{b},c):\ (\hat{\boldsymbol{b}},\hat{c})\in A, a<0\}} |a|(\|\boldsymbol{b}\|_1 + |c|)\rho(da, d\boldsymbol{b}, dc), \tag{4}$$

for any Borel set $A \subset \Lambda$, where

$$\hat{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_1 + |c|}, \quad \hat{c} = \frac{c}{\|\boldsymbol{b}\|_1 + |c|}. \tag{5}$$

Obviously $\rho_+(\Lambda) + \rho_-(\Lambda) = \mathbb{E}_\rho\left[|a|(\|\boldsymbol{b}\|_1 + |c|)\right]$, and

$$f(\boldsymbol{x}) = \int_\Lambda \sigma(\boldsymbol{b}^T\boldsymbol{x} + c)\rho_+(d\boldsymbol{b}, dc) - \int_\Lambda \sigma(\boldsymbol{b}^T\boldsymbol{x} + c)\rho_-(d\boldsymbol{b}, dc). \tag{6}$$

Next, we define extensions of ρ_+ and ρ_- to $\{-1, 1\} \times \Lambda$ by

$$\tilde{\rho}_+(A') = \rho_+(\{(\boldsymbol{b}, c) : \ (1, \boldsymbol{b}, c) \in A'\}), \tag{7}$$
$$\tilde{\rho}_-(A') = \rho_-(\{(\boldsymbol{b}, c) : \ (-1, \boldsymbol{b}, c) \in A'\}), \tag{8}$$

for any Borel sets $A' \subset \{-1, 1\} \times \Lambda$, and let $\tilde{\rho} = \tilde{\rho}_+ + \tilde{\rho}_-$. Then we have $\tilde{\rho}(\{-1, 1\} \times \Lambda) = \mathbb{E}_\rho\left[|a|(\|\boldsymbol{b}\|_1 + |c|)\right]$ and

$$f(\boldsymbol{x}) = \int_{\{-1,1\}\times\Lambda} a\sigma(\boldsymbol{b}^T\boldsymbol{x} + c)\tilde{\rho}(da, d\boldsymbol{b}, dc). \tag{9}$$

Therefore, we can normalize $\tilde{\rho}$ to be a probability measure, and

$$\|f\|_{\mathcal{B}_\infty} \leq \tilde{\rho}(\{-1, 1\} \times \Lambda) = \|f\|_{\mathcal{B}_1} < \infty. \tag{10}$$

Barron space as union of RKHS

Equivalent formulation of the integral representation:

$$f(\boldsymbol{x}) = \int a(\boldsymbol{w})\sigma(\boldsymbol{w}^T\tilde{\boldsymbol{x}})\pi(d\boldsymbol{w}), \quad \boldsymbol{w} = (\boldsymbol{b}, c),\ \tilde{\boldsymbol{x}} = (\boldsymbol{x}, 1)$$

Define:

$$k_\pi(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}'}) = \mathbb{E}_{\boldsymbol{w}\sim\pi}\sigma(\boldsymbol{w}^T\tilde{\boldsymbol{x}})\sigma(\boldsymbol{w}^T\tilde{\boldsymbol{x}'})$$

We have

$$\mathcal{B} = \bigcup_\pi \mathcal{H}_{k_\pi}$$

- Two-layer neural network can be understood as an "adaptive kernel method".
- The Barron space is larger than any RKHS, i.e. any kernel method suffers from CoD in the Barron space [Barron 93].

Direct approximation theorem

Define the path norm as

$$\|\theta\|_{\mathcal{P}} := \frac{1}{m}\sum_{j=1}^{m}|a_j|(\|\boldsymbol{b}_j\|_1 + |c_j|)$$

Theorem (Direct approximation theorem)

For any $f \in \mathcal{B}$ and integer $m > 0$, there exists a two-layer neural network f_m such that

$$\|f(\cdot) - f_m(\cdot;\theta)\|^2 \leq \frac{3\|f\|_{\mathcal{B}}^2}{m},$$

Furthermore, we have $\|\theta\|_{\mathcal{P}} \leq 2\|f\|_{\mathcal{B}}$.

Prove by i.i.d. sampling $\{(a_k, \boldsymbol{b}_k, c_k)\}_{i=1}^m$ from ρ and considering

$$\hat{f}_m(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m} a_k\sigma(\boldsymbol{b}_k^T\boldsymbol{x} + c_k).$$

Inverse approximation theorem

Define

$$\mathcal{N}_Q := \left\{ f_m(\boldsymbol{x};\theta) : \ \|\theta\|_{\mathcal{P}} \leq Q, m \in \mathbb{N}^+ \right\}.$$

Theorem (Inverse approximation theorem)

Let f be a continuous function on $[0,1]^d$. If there exists a constant Q and a sequence of functions $(f_m) \subset \mathcal{N}_Q$ such that

$$f_m(\boldsymbol{x}) \to f(\boldsymbol{x}) \ \ (m \to \infty)$$

for all $\boldsymbol{x} \in [0,1]^d$. Then $f \in \mathcal{B}$ and $\|f\|_{\mathcal{B}} \leq Q$.

Assume $f_m(\boldsymbol{x}) = \frac{1}{m} \sum_{k=1}^m a_k^{(m)} \sigma(\boldsymbol{b}_k^{(m)} \boldsymbol{x} + c_k^{(m)})$, prove the inverse approximation theorem by showing the sequence (ρ_m) with

$$\rho_m(a, \boldsymbol{b}, c) = \frac{1}{m} \sum_{k=1}^m \delta(a - a_k^{(m)}) \delta(\boldsymbol{b} - \boldsymbol{b}_k^{(m)}) \delta(c - c_k^{(m)})$$

is tight.

Implication from the approximation results

1. The Barron space catches all the functions that can be approximated by two-layer neural network with bounded path norm, and the approximation error does not suffer from CoD.

2. The Barron space is the right function set to study for two-layer neural networks, and the Barron norm is the natural norm associated with the Barron space.

3. (Next) Functions in the Barron space can be learned efficiently by two-layer neural networks.

Generalization error

Empirical risk: $\hat{\mathcal{R}}_{\mathcal{S}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(\boldsymbol{x}_i), y_i)$

Population risk: $\mathcal{R}(\theta) := \mathbb{E}_{\boldsymbol{x}\sim\mu^*} l(f_\theta(\boldsymbol{x}), y)$

- We are interested in $\mathcal{R}$, but we can only minimize $\hat{\mathcal{R}}_{\mathcal{S}}$. The difference between $\hat{\mathcal{R}}_{\mathcal{S}}$ and $\mathcal{R}$ (the generalization gap) may be large.
- Control the generalization gap: **Rademacher complexity**

Definition (Rademacher complexity)

Given a function class $\mathcal{H}$ and sample set $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$, the *(empirical) Rademacher complexity* of $\mathcal{H}$ with respect to $\mathcal{S}$ is defined as

$$\mathrm{Rad}_{\mathcal{S}}(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\xi\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\xi_i h(\boldsymbol{x}_i)\right],$$

where the ξ_i are independent random variables with $\Pr\{\xi_i = \pm 1\} = 1/2$.

Generalization error

Empirical risk: $\hat{\mathcal{R}}_{\mathcal{S}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(\boldsymbol{x}_i), y_i)$

Population risk: $\mathcal{R}(\theta) := \mathbb{E}_{\boldsymbol{x}\sim\mu^*} l(f_\theta(\boldsymbol{x}), y)$

- We are interested in $\mathcal{R}$, but we can only minimize $\hat{\mathcal{R}}_{\mathcal{S}}$. The difference between $\hat{\mathcal{R}}_{\mathcal{S}}$ and $\mathcal{R}$ (the generalization gap) may be large.
- Control the generalization gap: **Rademacher complexity**

Theorem

Given a function class $\mathcal{H}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random samples $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$,

$$\sup_{h\in\mathcal{H}} \left| \mathbb{E}_{\boldsymbol{x}} h(\boldsymbol{x}) - \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i) \right| \leq 2Rad_{\mathcal{S}}(\mathcal{H}) + 2 \sup_{h,h'\in\mathcal{H}} \|h - h'\|_\infty \sqrt{\frac{2\log(4/\delta)}{n}}$$

Generalization error

Norm-based generalization gap estimates:

- For model $f_\theta(\boldsymbol{x})$, design a parameter norm $\|\theta\|$
- Construct hypothesis spaces $\mathcal{H}^Q := \{f_\theta(\cdot) : \ \|\theta\| \leq Q\}$ and show

$$\mathrm{Rad}(\mathcal{H}^Q) \lesssim \frac{Q}{\sqrt{n}}$$

- Derive bounds for generalization gap by Rademacher complexity estimates

$$R(\theta) - \hat{R}_{\mathcal{S}}(\theta) \lesssim \frac{\|\theta\|}{\sqrt{n}}$$

  Path norm [Neyshabur et al.], Spectral norm [Bartlett et al.], Variational norm [Barron et al.]

- The bounds depend on training, and is usually vacuous.

Bounds for generalization errors

For finite element method (FEM), there are two types of error bounds:

$$\text{a posteriori estimates:} \qquad \|\hat{f}_n - f^*\|_1 \leq Cn^{-\alpha}\|\hat{f}_n\|_2,$$

$$\text{a priori estimates:} \qquad \|\hat{f}_n - f^*\|_1 \leq Cn^{-\alpha}\|f^*\|_3$$

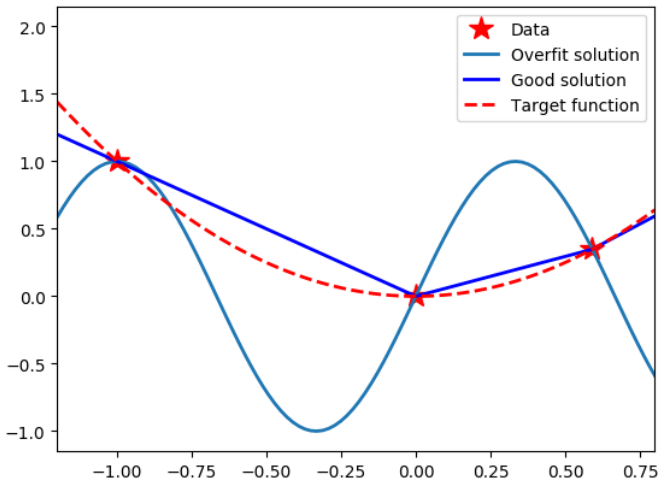
For neural networks, the norm-based bounds are a posteriori, since the bounds depend on parameters after training.

In this section, we derive a priori estimates for neural network models.

Towards a-priori estimates

To get a-priori estimates, we need:

- Regularization

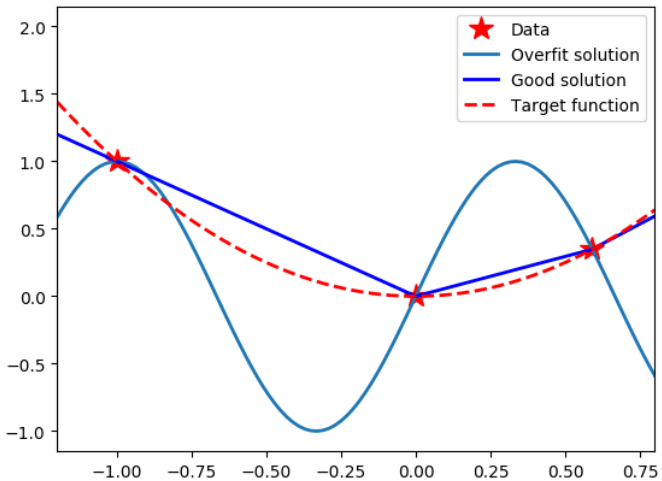


- A proper function space for target functions

Towards a-priori estimates

To get a-priori estimates, we need:

- Regularization ..... Path norm



- A proper function space for target functions ··· Barron space

Two-layer neural networks: main result

In the noiseless case ($y_i = f^*(\boldsymbol{x}_i)$), let

$$J_\lambda(\theta) := \hat{\mathcal{R}}_n(\theta) + \frac{\lambda}{\sqrt{n}}\|\theta\|_{\mathcal{P}},$$

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

Theorem

Assume that $\lambda \geq 4\sqrt{2\log(2d)}$, the width of the two-layer neural network is m. Then there exists a constant C, such that for any $\delta > 0$, with probability at least $1-\delta$ over the choice of the training set, we have

$$\mathcal{R}(\hat{\theta}_{n,\lambda}) \leq \underbrace{\frac{3\|f^*\|_{\mathcal{B}}^2}{m}}_{\text{Approximation Err.}} + \underbrace{\frac{C}{\sqrt{n}}\left((1+\lambda)\|f^*\|_{\mathcal{B}} + \sqrt{\log(14/\delta)}\right)}_{\text{Estimation Err.}}.$$

Proof sketch

Let $J(\theta) = \hat{\mathcal{R}}_n(\theta) + \frac{\lambda}{\sqrt{n}}\|\theta\|_{\mathcal{P}}$. Let $\tilde{\theta}$ be the solution given by the direct approximation theorem in the first section. We have

$$\mathcal{R}(\hat{\theta}_{n,\lambda}) = \underbrace{\mathcal{R}(\hat{\theta}_{n,\lambda}) - \hat{\mathcal{R}}_n(\hat{\theta}_{n,\lambda})}_{I} + \underbrace{\hat{\mathcal{R}}_n(\hat{\theta}_{n,\lambda}) - J(\hat{\theta}_{n,\lambda})}_{II} + \underbrace{J(\hat{\theta}_{n,\lambda}) - J(\tilde{\theta})}_{III}$$
$$+ \underbrace{J(\tilde{\theta}) - \hat{\mathcal{R}}_n(\tilde{\theta})}_{II} + \underbrace{\hat{\mathcal{R}}_n(\tilde{\theta}) - \mathcal{R}(\tilde{\theta})}_{I} + \mathcal{R}(\tilde{\theta}).$$

- $\mathcal{R}(\tilde{\theta})$**:** Approximation error.
- **I:** Estimation error (generalization gap).
- **II:** The regularization term, $\frac{\lambda}{\sqrt{n}}\|\theta\|_{\mathcal{P}}$.
- **III:** ≤ 0.

Proof sketch

The generalization gap is bounded by $\mathcal{O}(\frac{\|\theta\|_{\mathcal{P}}}{\sqrt{n}})$ because of the following bound of Rademacher complexity:

Theorem ([Neyshabur et al.])

Let $\mathcal{H}_Q = \{f(\boldsymbol{x};\theta) : \|\theta\|_{\mathcal{P}} \leq Q\}$. Then we have

$$\mathit{Rad}_n(\mathcal{H}_Q) \leq 2Q\sqrt{\frac{2\log 2d}{n}}.$$

The proof is completed by showing both $\|\hat{\theta}_{n,\lambda}\|_{\mathcal{P}}$ and $\|\tilde{\theta}\|_{\mathcal{P}}$ are small.

Result for noisy case

- Consider the noisy case:

$$y_i = f^*(\boldsymbol{x}_i) + \xi_i,$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. sub-Gaussian random variable, i.e.

$$\mathbb{P}[|\xi| > t] \leq c_0 e^{-\frac{t^2}{\sigma}} \quad \forall\, t \geq \tau_0.$$

- Consider the following regularized estimator:

$$\hat{\theta}_{n,\lambda} = \text{argmin}\, R_n(\theta) + \frac{\lambda B_n}{\sqrt{n}} \|\theta\|_{\mathcal{P}},$$

where $B_n = 1 + \max(\tau_0, \sigma^2 \ln(n))$.

Result for noisy case (cont'd)

Theorem (noisy case)

Assume that $\lambda \geq 4\sqrt{2\ln(2d)}$. *Then for any* $\delta > 0$, *with probability at least* $1-\delta$, *we have*

$$\mathbb{E}_x|f(x;\hat{\theta}_{n,\lambda}) - f^*(x)|^2 \lesssim \frac{\|f^*\|_{\mathcal{B}}^2}{m} + \frac{\lambda B_n}{\sqrt{n}}\|f^*\|_{\mathcal{B}}$$
$$+ \frac{B_n^2}{\sqrt{n}}\left(\|f^*\|_{\mathcal{B}} + \sqrt{\log(n/\delta)}\right) + \frac{B_n^2}{\sqrt{n}}\left(c_0\sigma^2 + \sqrt{\frac{\mathbb{E}[\xi^2]}{\lambda}}\right).$$

- The noise introduces only logarithmic terms.

Residual Networks: A simplified model

We consider a simplified residual network model whose residual blocks have only one non-linearity, defined by

$$\begin{aligned} \boldsymbol{z}_{0,L}(\boldsymbol{x}) &= \boldsymbol{V}\boldsymbol{x} \\ \boldsymbol{z}_{l+1,L}(\boldsymbol{x}) &= \boldsymbol{z}_{l,L}(\boldsymbol{x}) + \boldsymbol{U}_l \sigma \circ (\boldsymbol{W}_l z_{l,L}(\boldsymbol{x})) \\ f_L(\boldsymbol{x};\Theta) &= \boldsymbol{\alpha}^T \boldsymbol{z}_{L,L}(\boldsymbol{x}) \end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{V} \in \mathbb{R}^{D\times d}, \boldsymbol{W}_l \in \mathbb{R}^{m\times D}$, $\boldsymbol{U}_l \in \mathbb{R}^{D\times m}, \boldsymbol{\alpha} \in \mathbb{R}^D$. $\sigma(\cdot)$ is the ReLU function, and $\sigma \circ (A)$ for matrix A means applying σ to each element of A.

Let $\Theta := \{\boldsymbol{V}, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_L, \boldsymbol{W}_l, \ldots, \boldsymbol{W}_L, \boldsymbol{\alpha}\}$.

How to get a priori estimate

We build a priori estimates in a similar way as two-layer networks:

1. Find a parameter norm $\|\Theta\|$, and consider the hypothesis space

$$\mathcal{F}^Q := \{f(\boldsymbol{x};\Theta) : \ \|\Theta\| \leq Q\} \tag{11}$$

2. For target function f^* in the **Barron space**, show $\mathcal{F}^Q$ can well approximate f^*, with $Q \lesssim \|f^*\|_{\mathcal{B}}$
3. Control the Rademacher complexity of $\mathcal{F}^Q$, so as the estimation error.
4. Finally, consider the solution of the regularized minimizing problem

$$\hat{\Theta}_{n,\lambda} = \text{argmin } \hat{R}_n(\Theta) + \frac{\lambda}{\sqrt{n}}\|\Theta\|, \tag{12}$$

and derive a priori estimate for $\hat{\Theta}_{n,\lambda}$ by controlling approximation and estimation error simultaneously.

Key problem

What norm to use?

Path norm?

$$\|\Theta\|_{\mathcal{P}} = \sum_{\mathcal{P}:paths} \prod_{w\in\mathcal{P}} |w| \tag{13}$$

1. Good approximation

Approximation

For any target function $f^* \in \mathcal{B}$, and any $L, m \geq 1$, there exists a residual network $f(\cdot;\tilde{\Theta})$ with depth L and width m, such that

$$R(\tilde{\Theta}) \lesssim \frac{\|f^*\|_{\mathcal{B}}^2}{Lm}, \qquad \|\tilde{\theta}\|_{\mathcal{P}} \lesssim \|f^*\|_{\mathcal{B}}.$$

2. Bad complexity control

Rademacher comlexity

$$\mathrm{Rad}(\mathcal{F}^Q) \leq 2^L \sqrt{\frac{2\log(2d)}{n}}.$$

Spectral norm?

$$\|\Theta\|_{\mathrm{N}} = \left[\prod_{l=1}^{L} \|\boldsymbol{W}_l\|_\sigma\right] \left[\sum_{l=1}^{L} \frac{\|\boldsymbol{W}_l^T\|_{2,1}^{2/3}}{\|\boldsymbol{W}_l\|_\sigma^{2/3}}\right]^{3/2}, \qquad \text{(Bartlett et. al.)} \quad (14)$$

1. Good complexity control

Rademacher comlexity

$$\mathrm{Rad}(\mathcal{F}^Q) \leq 12 \log n \sqrt{\frac{2\log(2d)}{n}}.$$

2. Bad approximation

Approximation

For target function $f^* \in \mathcal{B}$, and any $L, m \geq 1$, there exists a residual network $f(\cdot; \tilde{\Theta})$ with depth L and width m, such that

$$R(\tilde{\Theta}) \lesssim \frac{\|f^*\|_{\mathcal{B}}^2}{Lm}, \qquad \|\tilde{\theta}\|_{\mathrm{N}} \lesssim (Lm)^{3/2} \|f^*\|_{\mathcal{B}}.$$

Balance of approximation and estimation

1. $\mathcal{F}^Q$ large so that the approximation error is small
2. $\mathcal{F}^Q$ small so that the estimation error is small

$\mathcal{F}^Q$ should have proper size to balance approximation and estimation errors

The weighted path norm

The Weighted path norm:

Definition (Weighted path norm)

$$\|\Theta\|_{\mathcal{P}} := \||\boldsymbol{\alpha}|^T(I + 2|\boldsymbol{U}_L||\boldsymbol{W}_L|)\cdots(I + 2|\boldsymbol{U}_1||\boldsymbol{W}_1|)|\boldsymbol{V}|\|_1$$

Another formulation:

$$\|\Theta\|_{\mathcal{P}} = \sum_{\mathcal{P}:paths} 2^{\mathrm{nl}(\mathcal{P})} \prod_{w\in\mathcal{P}} |w|, \tag{15}$$

where $\mathrm{nl}(\mathcal{P})$ is the number of non-linearities $\mathcal{P}$ goes through.

- Larger weights for paths going through more nonlinearities.

A priori estimate in Barron space

$$\hat{\theta}_{n,\lambda} = \arg\min_{\theta} \hat{\mathcal{R}}_n(\theta) + \frac{\lambda}{\sqrt{n}}\|\theta\|_{\mathrm{WP}}$$

Theorem

Assume that $D \geq d+1, \lambda \geq 4\sqrt{2\log(2d)} + 3$. *For any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *the population risk satisfies*

$$\mathcal{R}(\hat{\theta}_{n,\lambda}) \leq \frac{3\|f^*\|_{\mathcal{B}}^2}{Lm} + (4\|f^*\|_{\mathcal{B}} + 1)\frac{3(4+\lambda)\sqrt{2\log(2d)} + 2}{\sqrt{n}} + 4\sqrt{\frac{2\log(14/\delta)}{n}}.$$

- The bound scales near-optimally as $m, L, n \to \infty$.
- Only condition for width is that $D \geq d+1$.
- Can potentially be extended to the flow-induced function space

Proof: Approximation

Theorem

For any target function $f^* \in \mathcal{B}$, and any $L, m \geq 1$, there exists a residual network $f(\cdot; \tilde{\Theta})$ with depth L and width m, such that

$$R(\tilde{\theta}) \leq \frac{3\|f^*\|_{\mathcal{B}}^2}{Lm}, \qquad \|\tilde{\theta}\|_{WP} \leq 4\|f^*\|_{\mathcal{B}}.$$

Proof: Firstly, we can find a two-layer neural network

$$f(\boldsymbol{x}; \hat{\theta}) = \frac{1}{Lm} \sum_{k=1}^{Lm} a_k \sigma(\boldsymbol{b}_k^T \boldsymbol{x}),$$

such that

$$R(\hat{\theta}) \leq \frac{3\|f^*\|_{\mathcal{B}}^2}{Lm}, \quad \|\hat{\theta}\|_{\mathcal{P}} \leq 2\|f^*\|_{\mathcal{B}}.$$

Proof: Approximation

Then, we construct a residual network $\tilde{\theta}$ from the two layer network $\hat{\theta}$ by

$$\boldsymbol{V} = \begin{bmatrix} I_d & 0 \end{bmatrix}^T, \quad \alpha = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T,$$

$$\boldsymbol{W}_l = \begin{bmatrix} \boldsymbol{b}_{(l-1)m+1}^T & 0 \\ \boldsymbol{b}_{(l-1)m+2}^T & 0 \\ \vdots & \vdots \\ \boldsymbol{b}_{lm}^T & 0 \end{bmatrix}, \quad \boldsymbol{U}_l = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ a_{(l-1)m+1} & a_{(l-1)m+2} & \cdots & a_{lm} \end{bmatrix}$$

for $l = 1, \ldots, L$.

Split the two-layer network into L pieces and stack the pieces to form a deep residual network.

Proof: Generalization gap

Theorem

Let $\mathcal{H}^Q = \{f(\cdot;\theta) : \|\theta\|_{WP} \leq Q\}$ where $f(\cdot,\theta)$'s are residual networks. Then we have

$$Rad(\mathcal{H}^Q) \leq 2Q\sqrt{\frac{2\log(2d)}{n}}.$$

Proof: Let $\mathbf{g}_l(\mathbf{x}) = \sigma(\boldsymbol{W}_l \boldsymbol{z}_{l-1}), \quad l = 1, \ldots, L$, then we can define weighted path norm for $\mathbf{g}_l$. Let g_l be an element of $\mathbf{g}_l$, and define

$$\mathcal{G}_l^Q = \{g_l : \|g_l\|_{\mathrm{WP}} \leq Q\}$$

Then we prove the result by induction on $\mathrm{Rad}(\mathcal{G}_l^Q)$

Proof: Generalization gap

$$
\begin{aligned}
n\hat{R}(\mathcal{G}_{l+1}^{Q}) &= \mathbb{E}_{\xi} \sup_{g_{l+1}\in\mathcal{G}_{l+1}^{Q}} \sum_{i=1}^{n} \xi_i g_{l+1}(\mathbf{x}_i) \\
&= \mathbb{E}_{\xi} \sup_{(1)} \sum_{i=1}^{n} \xi_i \sigma(\boldsymbol{w}_{l+1}^{T}(\boldsymbol{U}_l \mathbf{g}_l + \boldsymbol{U}_{l-1}\mathbf{g}_{l-1} + \cdots + \boldsymbol{U}_1 \mathbf{g}_1 + \mathbf{h}_0)) \\
&\leq \mathbb{E}_{\xi} \sup_{(1)} \sum_{i=1}^{n} \xi_i (\boldsymbol{w}_{l+1}^{T}(\boldsymbol{U}_l \mathbf{g}_l + \boldsymbol{U}_{l-1}\mathbf{g}_{l-1} + \cdots + \boldsymbol{U}_1 \mathbf{g}_1 + \mathbf{h}_0)) \\
&= \mathbb{E}_{\xi} \sup_{(1)} \sum_{i=1}^{n} \xi_i (\boldsymbol{w}_{l+1}^{T}(\boldsymbol{U}_l \mathbf{g}_l + \boldsymbol{U}_{l-1}\mathbf{g}_{l-1} + \cdots + \boldsymbol{U}_1 \mathbf{g}_1 + \boldsymbol{V}\sigma(\mathbf{x}) - \boldsymbol{V}\sigma(-\mathbf{x}))) \\
&\leq \mathbb{E}_{\xi} \sup_{(2)} \left\{ \sum_{k=1}^{l} a_k \sup_{g\in\mathcal{G}_k^1} \left| \sum_{i=1}^{n} \xi_i g(\mathbf{x}_i) \right| + 2b \sup_{g\in\mathcal{G}_1^1} \left| \sum_{i=1}^{n} \xi_i g(\mathbf{x}_i) \right| \right\} \\
&\leq \mathbb{E}_{\xi} \sup_{\substack{2a+4b\leq Q \\ a,b\geq 0}} (a+2b) \sup_{g\in\mathcal{G}_l^1} \left| \sum_{i=1}^{n} \xi_i g(\mathbf{x}_i) \right| \\
&\leq \frac{Q}{2} \mathbb{E}_{\xi} \sup_{g\in\mathcal{G}_l^1} \left| \sum_{i=1}^{n} \xi_i g(\mathbf{x}_i) \right|
\end{aligned}
$$

Proof continued

Since $0 \in \mathcal{G}_l^1$, for any $\{\xi_1, \ldots, \xi_n\}$, we have

$$\sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n \xi_i g(\mathbf{x}_i) \geq 0.$$

Hence, we have

$$\begin{aligned}\sup_{g \in \mathcal{G}_l^1} \left| \sum_{i=1}^n \xi_i g(\mathbf{x}_i) \right| &\leq \max \left\{ \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n \xi_i g(\mathbf{x}_i), \ \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n -\xi_i g(\mathbf{x}_i) \right\} \\ &\leq \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n \xi_i g(\mathbf{x}_i) + \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n -\xi_i g(\mathbf{x}_i),\end{aligned}$$

which gives

$$\mathbb{E}_\xi \sup_{g \in \mathcal{G}_l^1} \left| \sum_{i=1}^n \xi_i g(\mathbf{x}_i) \right| \leq 2\mathbb{E}_\xi \sup_{g \in \mathcal{G}_l^1} \sum_{i=1}^n \xi_i g(\mathbf{x}_i) = 2n\hat{R}(\mathcal{G}_l^1).$$

Therefore, we have

$$\hat{R}(\mathcal{G}_{l+1}^Q) \leq \frac{Q}{2} 2\sqrt{\frac{2\log(2d)}{n}} \leq Q\sqrt{\frac{2\log(2d)}{n}}.$$

Noisy case

- Consider the noisy case $y_i = f^*(\boldsymbol{x}_i) + \xi_i$, where $\{\xi_i\}_{i=1}^n$ are i.i.d. sub-Gaussian random variable, i.e.

$$\mathbb{P}[|\xi| > t] \leq c_0 e^{-\frac{t^2}{\sigma}} \quad \forall\, t \geq \tau_0.$$

- Let $B_n = 1 + \max(\tau_0, \sigma^2 \log(n))$, and

$$\hat{\theta}_{n,\lambda} = \arg\min_{\theta} \mathcal{R}_n(\theta) + \frac{\lambda B_n}{\sqrt{n}} \|\theta\|_{\mathrm{WP}},$$

Theorem (noisy case)

Assume that $\lambda \geq 4\sqrt{2\log(2d)} + 3B_n$. *Then, for any* $\delta > 0$, *with probability at least* $1 - \delta$, *we have*

$$R(\hat{\theta}_{n,\lambda}) \leq \frac{3\|f^*\|_{\mathcal{B}}^2}{Lm} + (6\|f^*\|_{\mathcal{B}} + 1)\frac{3(4+\lambda)B_n\sqrt{2\log(2d)} + 2B_n^2}{\sqrt{n}}$$
$$+ 4B_n^2\sqrt{\frac{2\log(14/\delta)}{n}} + \frac{2c(4\sigma^2+1)}{\sqrt{n}}.$$

Summary

- While the norm-based a posteriori estimates only requires the hypothesis space controlled by the parameter norm $\mathcal{H}^Q$ to have low complexity, the a priori estimates requires tradeoff between approximation ability and complexity.
- If $\mathcal{H}^Q$ is too small, it has small estimation error but large approximation error. If $\mathcal{H}^Q$ is too large, it has small approximation error but large estimation error. In this work, we design parameter norms so that $\mathcal{H}^Q$ has the right size, and hence strike a balance between approximation and estimation.
- Similar analysis can be applied to noisy case, classification problems, and other models such as autoencoders. It can also be extended to neural networks with general activation function other than ReLU.

Table of Contents

Residual networks: the flow-induced function space

Residual networks revisited:

$$\begin{aligned} \boldsymbol{z}_{0,L}(\boldsymbol{x}) &= \boldsymbol{V}\boldsymbol{x}, \\ \boldsymbol{z}_{l+1,L}(\boldsymbol{x}) &= \boldsymbol{z}_{l,L}(\boldsymbol{x}) + \frac{1}{L}\boldsymbol{U}_l \sigma \circ (\boldsymbol{W}_l z_{l,L}(\boldsymbol{x})), \\ f_L(\boldsymbol{x};\theta) &= \boldsymbol{\alpha}^T \boldsymbol{z}_{L,L}(\boldsymbol{x}) \end{aligned}$$

where $\boldsymbol{V} = [I_d, 0]^T \in \mathbb{R}^{D\times d}$, $\boldsymbol{z}_{l,L} \in \mathbb{R}^D$.

Flow-based continuous formulation:

$$\begin{aligned} \boldsymbol{z}(\boldsymbol{x},0) &= \boldsymbol{V}\boldsymbol{x}, \\ \dot{\boldsymbol{z}}(\boldsymbol{x},t) &= \mathbb{E}_{(\boldsymbol{U},\boldsymbol{W})\sim\rho_t} \boldsymbol{U}\sigma(\boldsymbol{W}\boldsymbol{z}(\boldsymbol{x},t)), \\ f_{\boldsymbol{\alpha},\{\rho_t\}}(\boldsymbol{x}) &= \boldsymbol{\alpha}^T \boldsymbol{z}(\boldsymbol{x},1), \end{aligned}$$

The flow-induced function space

Let

$$N_p(0) = \boldsymbol{e},$$
$$\dot{N}_p(t) = \left(\mathbb{E}_{\rho_t}(|\boldsymbol{U}||\boldsymbol{W}|)^p\right)^{1/p} N_p(t),$$

we define the flow-induced norm as

$$\|f\|_{\mathcal{D}_p} = \inf_{f=f_{\boldsymbol{\alpha},\{\rho_t\}}} |\boldsymbol{\alpha}|^T N_p(1) + \|N_p(1)\|_1 - D,$$

and the flow-induced function space

$$\mathcal{D}_p := \{f : \ \|f\|_{\mathcal{D}_p} < \infty\}$$

Path norm for residual networks:

$$\|\theta\|_{\mathcal{P}} = |\boldsymbol{\alpha}|^T \prod_{l=1}^{L} \left(I + \frac{1}{L}|\boldsymbol{U}_l||\boldsymbol{W}_l|\right) \boldsymbol{e}.$$

Direct and approximation theorems

Theorem (Direct approximation theorem (informal))

Let $f \in \mathcal{D}_1$, $\delta \in (0,1)$. If $f = f_{\boldsymbol{\alpha},\{\rho_t\}}$ and $\{\rho_t\}$ satisfies the "Lipschitz condition" on t. Then, there exists an L_0, which depends polynomially on $\|f\|_{\mathcal{D}_1}$ and δ, such that for any $L \geq L_0$, there exists an L-layer residual network $f_L(\cdot;\theta)$ that satisfies

$$\|f - f_L(\cdot;\theta)\|^2 \leq \frac{3\|f\|_{\mathcal{D}_1}^2}{L^{1-\delta}},$$

and $\|\theta\|_{\mathcal{P}} \leq 9\|f\|_{\mathcal{D}_1}$

Prove by the "compositionoal law of large numbers", i.e. if the weights $(\boldsymbol{U}_l, \boldsymbol{W}_l)$ for a residual network are independently sampled from $\rho_{l/L}$, for $l = 0, 1, ..., L-1$, then

$$\boldsymbol{z}_{L,L}(\boldsymbol{x}) \xrightarrow{p} \boldsymbol{z}(\boldsymbol{x}, 1), \quad \text{uniformly in } \boldsymbol{x}.$$

Inverse approximation theorem

Theorem (Inverse approximation theorem (informal))

Let f be a function defined on $[0,1]^d$. Assume that there is a sequence of residual networks $\{f_L(\cdot;\theta_L)\}_{L=1}^{\infty}$ with uniformly bounded parameters such that $\|f(\boldsymbol{x}) - f_L(\boldsymbol{x};\theta)\| \to 0$. Then, we have

$$f \in \mathcal{D}_\infty \subset \mathcal{D}_1$$

Moreover, if there exists B s.t. $\|\theta_L\|_{\mathcal{P}} \leq B$, then $\|f\|_{\mathcal{D}_1} \leq CB$ for some constant C.

Prove by the theorem of Young measure: let $\boldsymbol{U}_t^L$ and $\boldsymbol{W}_t^L$ be the piece-wise constant interpolations of $(\boldsymbol{U}_l^L)$ and $(\boldsymbol{W}_l^L)$ in $t \in [0,1]$, respectively. Then, there exists a subsequence $\{L_k\}$ and a family of probability measure $\{\rho_t, t \in [0,1]\}$, such that for every Caratheodory function F,

$$\lim_{k\to\infty} \int_0^1 F(\boldsymbol{U}_t^{L_k}, \boldsymbol{W}_t^{L_k}, t)dt = \int_0^1 \mathbb{E}_{\rho_t} F(\boldsymbol{U}, \boldsymbol{W}, t)dt.$$

Comparison with Barron space

- Barron functions belong to the flow-induced function space. For any function $f \in \mathcal{B}$, and $D \geq d+2$ and $m \geq 1$, we have $f \in \mathcal{D}_1$, and

$$\|f\|_{\mathcal{D}_1} \leq 2\|f\|_{\mathcal{B}} + 1.$$

- Composition of Barron functions belongs to the flow-induced function space. For any $d > 0$, let $\mathcal{B}_d$ be the Barron space on $[0,1]^d$. Let $g \in \mathcal{B}_d$, $h \in \mathcal{B}_1$, and assume that the range of g lies in $[0,1]$. Let $f = h \circ g$ be the composition of g and h. Then we have $f \in \mathcal{D}_1$ and

$$\|f\|_{\mathcal{D}_1} \leq (\|h\|_{\mathcal{B}_1} + 1)(\|g\|_{\mathcal{B}_d} + 1).$$

The "depth separation"

Table of Contents

Summary

- We build function space theories for neural network models, and define Barron space for two-layer neural networks and flow-induced function space for residual networks
- TThe direct and inverse theorems show that the function spaces contain all the functions that can be approximated by well-behaved neural networks (without the curse of dimensionality), and the norms control the constant factors. The norms can also control the estimation error.
- The comparison between RKHS, Barron space and compositional space show the advantage of residual networks over two-layer networks, and two-layer networks over kernel method.

What's Next?

- Function spaces for other architectures, even for general architectures.
- Models $\rightarrow$ Spaces, Spaces $\rightarrow$ Models?
- Optimization issues: can algorithms explore the whole space, and can they find generalizable solution?