Lecture 14: DNN Generalization Theory

Haizhao Yang
Department of Mathematics
University of Maryland College Park

2022 Summer Mini Course
Tianyuan Mathematical Center in Central China

Supervised deep learning

Conditions

- Given data pairs $\{(x_i, y_i = f(x_i))\}$ from an unknown map $f(x)$ defined on Ω
- $\{x_i\}_{i=1}^n$ are sampled randomly from an unknown distribution $U(x)$ on Ω

Goal

Recover the unknown map $f(x)$

Deep learning in practice

- Only the empirical loss is available:

$$R_S(\theta) := \frac{1}{N}\sum_{i=1}^{N}(h(x_i;\theta) - y_i)^2$$

- The best empirical solution is $h(x;\theta_S)$ with

$$\theta_S = \text{argmin}\, R_S(\theta)$$

- Numerical optimization to obtain a numerical solution $h(x;\theta_N)$.
- In practice, $\theta_N \neq \theta_S$ and how good θ_N is?

Supervised machine learning

How large is the actual prediction error $R_D(\theta_N)$?

$$\begin{aligned} R_D(\theta_N) &= [R_D(\theta_N) - R_S(\theta_N)] + [R_S(\theta_N) - R_S(\theta_S)] + [R_S(\theta_S) - R_S(\theta_D)] \\ &\quad + [R_S(\theta_D) - R_D(\theta_D)] + R_D(\theta_D) \\ &\le R_D(\theta_D) + [R_S(\theta_N) - R_S(\theta_S)] \\ &\quad + [R_D(\theta_N) - R_S(\theta_N)] + [R_S(\theta_D) - R_D(\theta_D)], \end{aligned}$$

- $R_D(\theta_D) = \int_\Omega (h(x;\theta_D) - f(x))^2 d\mu(x) \le= \int_\Omega (h(x;\tilde{\theta}) - f(x))^2 d\mu(x)$ can be bounded by a constructive approximation of $\tilde{\theta}$
- $[R_S(\theta_N) - R_S(\theta_S)]$ is the optimization error
- Other two terms are the generalization error

This lecture discusses the case when $h(x;\theta)$ is a deep neural network.

Generalization of PDE solvers

Neural networks + least square for PDEs (date back to 1990s),

$$\mathcal{D}(u) = f \quad \text{in } \Omega,$$
$$\mathcal{B}(u) = g \quad \text{on } \partial\Omega.$$

A DNN $\phi(\boldsymbol{x};\boldsymbol{\theta}^*)$ is constructed to approximate the solution $u(\boldsymbol{x})$ via

$$\begin{aligned}
\boldsymbol{\theta}_D &:= \operatorname*{argmin}_{\boldsymbol{\theta}} R_D(\boldsymbol{\theta}) \\
&:= \operatorname*{argmin}_{\boldsymbol{\theta}} \|\mathcal{D}\phi(\boldsymbol{x};\boldsymbol{\theta}) - f(\boldsymbol{x})\|_2^2 + \lambda\|\mathcal{B}\phi(\boldsymbol{x};\boldsymbol{\theta}) - g(\boldsymbol{x})\|_2^2
\end{aligned}$$

or

$$\begin{aligned}
\boldsymbol{\theta}_D &:= \operatorname*{argmin}_{\boldsymbol{\theta}} R_D(\boldsymbol{\theta}) \\
&:= \operatorname*{argmin}_{\boldsymbol{\theta}} \|\mathcal{D}\phi(\boldsymbol{x};\boldsymbol{\theta}) - f(\boldsymbol{x})\|_2^2
\end{aligned}$$

if the DNN satisfies the boundary condition automatically.

Generalization of PDE solvers

For simplicity, consider:

$$\begin{aligned}\boldsymbol{\theta}_D &:= \operatorname*{argmin}_{\boldsymbol{\theta}} R_D(\boldsymbol{\theta}) \\ &:= \operatorname*{argmin}_{\boldsymbol{\theta}} \|\mathcal{D}\phi(\boldsymbol{x};\boldsymbol{\theta}) - f(\boldsymbol{x})\|_2^2.\end{aligned}$$

Discretization:

$$\begin{aligned}\boldsymbol{\theta}_S = \operatorname*{argmin}_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}) &:= \frac{1}{n}\sum_{S=\{\boldsymbol{x}_i\}_{i=1}^n \subset \Omega} \ell(\mathcal{L}\phi(\boldsymbol{x}_i;\boldsymbol{\theta}), f(\boldsymbol{x}_i)) \\ &:= \frac{1}{2n}\sum_{i=1}^{n}(\mathcal{L}\phi(\boldsymbol{x}_i;\boldsymbol{\theta}) - f(\boldsymbol{x}_i))^2\end{aligned}$$

Analysis goal: $R_D(\boldsymbol{\theta}_S) \leq ?$

Generalization of PDE solvers

What do we care?

Dimension independent rate of the generalization error.

- Low-dimensional mainifold assumption (arXiv:2104.06708)
- Low-complexity assumption (arXiv:1810.06397,arXiv:1908.11140)

Let us focus on the second case for PDE problems to show $R_D(\boldsymbol{\theta}_S) \leq O(\frac{1}{\sqrt{n}})$.

Generalization of PDE solvers

Functions with low complexity:

Definition (Barron Type Function)

A function $f : \Omega \to \mathbb{R}$ is called a Barron-type function if f has an integral representation

$$f(\boldsymbol{x}) = \mathbb{E}_{(a,\boldsymbol{w})\sim\rho} a[\boldsymbol{w}^\intercal \boldsymbol{A}(\boldsymbol{x})\boldsymbol{w}\sigma''(\boldsymbol{w}^\intercal \boldsymbol{x}) + \boldsymbol{b}^\intercal(\boldsymbol{x})\boldsymbol{w}\sigma'(\boldsymbol{w}^\intercal \boldsymbol{x}) + c(\boldsymbol{x})\sigma(\boldsymbol{w}^\intercal \boldsymbol{x})],$$

where ρ is a probability distribution over $\mathbb{R}^{d+1}$.

Definition (Barron Norm)

The associated Barron norm of a Barron-type function f is defined as

$$\|f\|_{\mathcal{B}} := \inf_{\rho \in \mathcal{P}_f} \left(\mathbb{E}_{(a,\boldsymbol{w})\sim\rho} |a|^2 \|\boldsymbol{w}\|_1^6\right)^{1/2},$$

where $\mathcal{P}_f = \{\rho \mid f(\boldsymbol{x}) = \mathbb{E}_{(a,\boldsymbol{w})\sim\rho} a[\boldsymbol{w}^\intercal \boldsymbol{A}(\boldsymbol{x})\boldsymbol{w}\sigma''(\boldsymbol{w}^\intercal \boldsymbol{x}) + \boldsymbol{b}^\intercal(\boldsymbol{x})\boldsymbol{w}\sigma'(\boldsymbol{w}^\intercal \boldsymbol{x}) + c(\boldsymbol{x})\sigma(\boldsymbol{w}^\intercal \boldsymbol{x})], \boldsymbol{x} \in \Omega\}$.

Definition (Barron Space)

The Barron-type space is defined as

$$\mathcal{B}(\Omega) = \{f : \Omega \to \mathbb{R} \mid \|f\|_{\mathcal{B}} < \infty\}.$$

Generalization of PDE solvers

Neural networks to be used to parameterize PDE solutions:

Definition (Path norm)

The path norm of a two-layer neural network

$$\phi(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{N} a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}),$$

with an activation function σ and a parameter set $\boldsymbol{\theta}$ is defined as

$$\|\boldsymbol{\theta}\|_{\mathcal{P}} := \sum_{j=1}^{N} |a_j| \|\boldsymbol{w}_j\|_1^3.$$

Consider $\sigma(x) = \max\{\frac{1}{6}x^3, 0\}$.

Generalization of PDE solvers

Consider the second order differential operator $\mathcal{L}$:

$$\mathcal{L}u = \sum_{\alpha,\beta=1}^{d} A_{\alpha\beta}(\boldsymbol{x})u_{x_\alpha x_\beta} + \sum_{\alpha=1}^{d} b_\alpha(\boldsymbol{x})u_{x_\alpha} + c(\boldsymbol{x})u.$$

Assumption (Symmetry and boundedness)

Assume $\mathcal{L}$ satisfies the condition: there exists $M \geq 1$[1] such that for all $\boldsymbol{x} \in \Omega = [0,1]^d$, $\alpha, \beta \in [d]$, we have $A_{\alpha\beta} = A_{\beta\alpha}$

$$|A_{\alpha\beta}(\boldsymbol{x})| \leq M, \quad |b_\alpha(\boldsymbol{x})| \leq M, \quad \textit{and} \quad |c(\boldsymbol{x})| \leq M.$$

[1] The upper bound M is not necessarily greater than 1. We set this for simplicity.

Generalization of PDE solvers

Luo and Y., arXiv:2006.15733

Theorem (A posteriori generalization bound)

For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of random sample locations $S := \{\boldsymbol{x}_i\}_{i=1}^n$, for any two-layer neural network $\phi(\boldsymbol{x};\boldsymbol{\theta})$, we have[2]

$$|R_{\mathcal{D}}(\boldsymbol{\theta}) - R_S(\boldsymbol{\theta})| \quad \leq \quad O\left(\frac{(\|\boldsymbol{\theta}\|_{\mathcal{P}}+1)^2}{\sqrt{n}} d^2\right).$$

Observation:

- This is the difference of the average of n samples and the true expectation: $\leq \frac{?}{\sqrt{n}}$
- This bound works for all possible two-layer NNs:
  - Bad NNs –> larger bounds
  - Good NNs –> smaller bounds
- This bound is $\frac{\text{Complexity of NNs}}{\sqrt{n}}$

[2] Ignoring prefactors and log terms.

Definition (The Rademacher complexity of a function class $\mathcal{F}$)

Given a sample set $S = \{z_1, \ldots, z_n\}$ on a domain $\mathcal{Z}$, and a class $\mathcal{F}$ of real-valued functions defined on $\mathcal{Z}$, the empirical Rademacher complexity of $\mathcal{F}$ on S is defined as

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\tau}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\tau_i f(z_i)\right],$$

where $\tau_1, \ldots, \tau_n$ are independent random variables drawn from the Rademacher distribution, i.e., $\mathbb{P}(\tau_i = +1) = \mathbb{P}(\tau_i = -1) = \frac{1}{2}$ for $i = 1, \ldots, n$.

Complexity of NNs

How to estimate the Rademacher complexity of NNs?

- $NN(x; \mathbf{a}, \mathbf{w}) = \mathbf{a}^{\mathbf{T}} \boldsymbol{\sigma}(\mathbf{wx})$: a linear transform of $\boldsymbol{\sigma}(\mathbf{w}x)$
- $\boldsymbol{\sigma}(\mathbf{w}x) = (\sigma(\boldsymbol{w}_1 x), \ldots, \sigma(\boldsymbol{w}_N x))$: the composition of $\boldsymbol{\sigma}$ and linear transforms of x
- Hence, NNs are the composition of a linear transform, σ, and linear transforms of x

Basic Rademacher complexity

- Function compositions
- Linear transformation

Lemma (Contraction lemma[3])

Suppose that $\psi_i : \mathbb{R} \to \mathbb{R}$ *is a* C_{L}*-Lipschitz function for each* $i \in [n]$*. For any* $\boldsymbol{y} \in \mathbb{R}^n$*, let* $\boldsymbol{\psi}(\boldsymbol{y}) = (\psi_1(y_1), \cdots, \psi_n(y_n))^\intercal$*. For an arbitrary set of vector functions* $\mathcal{F}$ *of length* n *on an arbitrary domain* $\mathcal{Z}$ *and an arbitrary choice of samples* $S = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\} \subset \mathcal{Z}$*, we have*

$$\mathrm{Rad}_S(\boldsymbol{\psi} \circ \mathcal{F}) \leq C_{\mathrm{L}} \mathrm{Rad}_S(\mathcal{F}).$$

[3] Understanding machine learning: From theory to algorithms, Shalev-Shwartz, S. and Ben-David, S.

Complexity of NNs

Lemma (Rademacher complexity for linear predictors[4])

Let $\Theta = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N\} \in \mathbb{R}^d$. *Let* $\mathcal{G} = \{g(\boldsymbol{w}) = \boldsymbol{w}^\intercal \boldsymbol{x} : \|\boldsymbol{x}\|_1 \le 1\}$ *be the linear function class with parameter* $\boldsymbol{x}$ *whose* ℓ^1 *norm is bounded by* 1*. Then*

$$\mathrm{Rad}_\Theta(\mathcal{G}) \le \max_{1\le k\le m} \|\boldsymbol{w}_k\|_\infty \sqrt{\frac{2\log(2d)}{N}}.$$

[4] Understanding machine learning: From theory to algorithms, Shalev-Shwartz, S. and Ben-David, S.

Complexity of NNs

Let us state a general theorem concerning the Rademacher complexity and generalization gap of an arbitrary set of functions $\mathcal{F}$ on an arbitrary domain $\mathcal{Z}$.

Theorem (Rademacher complexity and generalization gap[5])

Suppose that f's in $\mathcal{F}$ are non-negative and uniformly bounded, i.e., for any $f \in \mathcal{F}$ and any $\boldsymbol{z} \in \mathcal{Z}$, $0 \le f(\boldsymbol{z}) \le B$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of n i.i.d. random samples $S = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\} \subset \mathcal{Z}$, we have

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{z}_i) - \mathbb{E}_{\boldsymbol{z}} f(\boldsymbol{z}) \right| \le 2\mathbb{E}_S \mathrm{Rad}_S(\mathcal{F}) + B\sqrt{\frac{\log(2/\delta)}{2n}},$$

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{z}_i) - \mathbb{E}_{\boldsymbol{z}} f(\boldsymbol{z}) \right| \le 2\mathrm{Rad}_S(\mathcal{F}) + 3B\sqrt{\frac{\log(4/\delta)}{2n}}.$$

---

[5] Understanding machine learning: From theory to algorithms, Shalev-Shwartz, S. and Ben-David, S.

Generalization of PDE solvers

Luo and Y., arXiv:2006.15733

Theorem (A posteriori generalization bound)

For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of random sample locations $S := \{\boldsymbol{x}_i\}_{i=1}^n$, for any two-layer neural network $\phi(\boldsymbol{x};\boldsymbol{\theta})$, we have[6]

$$|R_{\mathcal{D}}(\boldsymbol{\theta}) - R_S(\boldsymbol{\theta})| \quad \leq \quad O\left(\frac{(\|\boldsymbol{\theta}\|_{\mathcal{P}}+1)^2}{\sqrt{n}} d^2\right).$$

Proof:

- $|R_{\mathcal{D}}(\boldsymbol{\theta}) - R_S(\boldsymbol{\theta})| \leq$ Rademacher complexity + Stat error $\leq O\left(\frac{\|\boldsymbol{\theta}\|_{\mathcal{P}}}{\sqrt{n}}\right) + O\left(\frac{1}{\sqrt{n}}\right)$.
- Apply the previous theorem with $f(x) = |NN(x;\theta) - y|^2$, which is the composition of $|x-y|^2$ and $NN(x;\theta)$.
- The Rademacher complexity of f is reduced to the one of NNs

[6] Ignoring prefactors and log terms.

Hard constraint as regularization

Corollary

Suppose that $f(\boldsymbol{x})$ is in the Barron-type space $\mathcal{B}([0,1]^d)$ and let

$$\boldsymbol{\theta}_{S,B} = \underset{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|_\infty \leq B}{\operatorname{argmin}} R_S(\boldsymbol{\theta}).$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of random samples $S := \{\boldsymbol{x}_i\}_{i=1}^n$, we have

$$\begin{aligned} R_{\mathcal{D}}(\boldsymbol{\theta}_{S,B}) &:= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \frac{1}{2}(\mathcal{L}\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,B}) - f(\boldsymbol{x}))^2 \\ &\leq R_S(\boldsymbol{\theta}_{S,B}) + |R_{\mathcal{D}}(\boldsymbol{\theta}_{S,B}) - R_S(\boldsymbol{\theta}_{S,B})| \\ &\leq O\left(\frac{\|f\|_{\mathcal{B}}^2 C_f^4}{N \min\{C_f^4, B^4\}}\right) + O\left(\frac{B^8 N^2 d^2}{\sqrt{n}}\right). \end{aligned}$$

Generalization of PDE solvers

Regression: E, Ma, and Wu, CMS, 2019
PDE solvers: Luo and Y., arXiv:2006.15733
Soft constraint as regularization

Theorem (A priori generalization bound)

Suppose that $f(\boldsymbol{x})$ is in the Barron-type space $\mathcal{B}([0,1]^d)$ and $\lambda \geq 4M^2[2 + 14d^2\sqrt{2\log(2d)} + \sqrt{2\log(2/3\delta)}]$. Let

$$\boldsymbol{\theta}_{S,\lambda} = \arg\min_{\boldsymbol{\theta}} J_{S,\lambda}(\boldsymbol{\theta}) := R_S(\boldsymbol{\theta}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{\theta}\|_{\mathcal{P}}^2 \log[\pi(\|\boldsymbol{\theta}\|_{\mathcal{P}} + 1)].$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of random samples $S := \{\boldsymbol{x}_i\}_{i=1}^n$, we have

$$\begin{aligned} R_{\mathcal{D}}(\boldsymbol{\theta}_{S,\lambda}) &:= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\frac{1}{2}(\mathcal{L}\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,\lambda}) - f(\boldsymbol{x}))^2 \\ &\leq O\left(\frac{\|f\|_{\mathcal{B}}^2}{N}\right) + O\left(\frac{\|f\|_{\mathcal{B}}^2}{\sqrt{n}}\right). \end{aligned}$$

Proof: $R_{\mathcal{D}}(\boldsymbol{\theta}_{S,\lambda}) \leq$ Approximation error + Rademacher complexity + Stat error $\leq O\left(\frac{\|f\|_{\mathcal{B}}^2}{N}\right) + O\left(\frac{\|\boldsymbol{\theta}\|_{\mathcal{P}}}{\sqrt{n}}\right) + O\left(\frac{1}{\sqrt{n}}\right) \leq O\left(\frac{\|f\|_{\mathcal{B}}^2}{N}\right) + O\left(\frac{\|f\|_{\mathcal{B}}^2}{\sqrt{n}}\right)$.