The Mathematical Theory of Neural Network-Based Machine Learning

Prof. Weinan E
Department of Mathematics
Princeton University

Typed by Yuelin Li

2019 APMA colloquium

1 Approximating Functions Using Samples

Given a set of samples from $\mu, \{\boldsymbol{x}_j\}_{j=1}^n$, and $\{y_j = f^*(\boldsymbol{x}_j)\}_{j=1}^n$
Approximate f^* using $S = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^n$.
Strategy:
Define some "hypothesis space" (set of functions) $\mathcal{H}_m$ ($m \sim$ dim of $\mathcal{H}_m$) :

$$\mathcal{H}_m = \{f(\cdot, \theta)\}$$

Best approximation minimizes the "population risk":

$$\mathcal{R}(\theta) = \mathbb{E}\left(f(\boldsymbol{x}, \theta) - f^*(\boldsymbol{x})\right)^2 = \int_{\mathbb{R}^d} \left(f(\boldsymbol{x}, \theta) - f^*(\boldsymbol{x})\right)^2 d\mu$$

In practice: Minimize the "empirical risk":

$$\mathcal{R}_n(\theta) = \frac{1}{n}\sum_j \left(f\left(\boldsymbol{x}_j, \theta\right) - y_j\right)^2 = \frac{1}{n}\sum_j \left(f\left(\boldsymbol{x}_j, \theta\right) - f^*\left(\boldsymbol{x}_j\right)\right)^2$$

To choose the hypothesis space:
$f(\boldsymbol{x}) = \beta \cdot \boldsymbol{x} + \beta_0$. Generalized linear models (feature-based models): $f(\boldsymbol{x}) = \sum_{k=1}^m c_k \phi_k(\boldsymbol{x})$, where $\{\phi_k\}$ are linearly independent functions.
Simple neural networks: $f(\boldsymbol{x}) = \sum_k a_k \sigma\left(\boldsymbol{b}_k \cdot \boldsymbol{x} + c_k\right)$, where σ is some nonlinear function.
- $\sigma(x) = \max(x, 0)$, the ReLU (rectified linear units) function.
- $\sigma(x) = (1 + e^{-x})^{-1}$.
- $\sigma(x) = \cos(x)$
- $\sigma(x) = \tanh(x)$

Deep neural networks (DNN): compositions of functions of the form above. (σ is a scalar function)

2 Dynamical System Viewpoint to Deep Learning

Constructing nonlinear approximations through the flow map of a dynamical system:

$$\frac{d\boldsymbol{z}(\boldsymbol{x},t)}{dt} = \boldsymbol{F}(\boldsymbol{z}(\boldsymbol{x},t)), \quad \boldsymbol{z}(0,\boldsymbol{x}) = \boldsymbol{V}\boldsymbol{x}$$

The flow map $x \to z(x,1)$ is a nonlinear mapping. Simplest choice of nonlinear $\boldsymbol{F}$:

$$\boldsymbol{F}(\boldsymbol{z};\boldsymbol{U},\boldsymbol{W}) = \boldsymbol{U}\sigma \circ (\boldsymbol{W}\boldsymbol{z})$$

Choose the optimal $\boldsymbol{U},\boldsymbol{W}(\cdot),\alpha$ to approximate f^* by

$$f^*(\boldsymbol{x}) \sim \alpha \cdot \boldsymbol{z}(\boldsymbol{x},1)$$

Application: High-dimensional control and PDEs high-dimensional stochastic control, nonlinear parabolic PDE...
Application: Molecular dynamics traditional dilemma cost VS accuracy.

$$E = E\left(x_1, x_2, \ldots, x_i, \ldots, x_N\right)$$
$$m_i \tfrac{d^2 x_i}{dt^2} = F_i = -\nabla_{x_i} E$$

Two ways to calculate E and F: 1) Empirical potentials: efficient but unreliable. The Lennard-Jones potential:

$$V_{ij} = 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right], E = \frac{1}{2}\sum_{i\neq j} V_{ij}$$

2) Multi-scale modeling: computing the inter-atomic forces on the fly using QM, e.g. the Car-Parrinello MD. Accurate but expensive:

$$E = \left\langle \Psi_0 \left| H_e^{KS} \right| \Psi_0 \right\rangle, \mu\ddot{\phi}_i = H_e^{KS}\phi_i + \sum_j \Lambda_{ij}\phi_j$$

Issues: Basic questions we want to understand: high dimensionality; models are highly over-parametrized, classical machine learning theory would suggest overfitting; models are non-convex, yet simple gradient algorithms seem to work (compare with structural optimization in material science and chemistry)
More advanced questions: why deep networks seem to perform better than shallow ones? Why stochastic gradient descent seems to perform better than gradient descent? etc.

3 Classical Numerical Analysis

We define a well-posed math model (the hypothesis space, the loss function...)
splines: hypothesis space = C^1 piecewise cubic polynomials the data

$$I_n(f) = \frac{1}{n}\sum_{j=1}^{n}\left(f\left(x_j\right) - y_j\right)^2 + \lambda\int\left|D^2 f(x)\right|^2 dx$$

finite elements: hypothesis space = C^0 piecewise polynomials.

We identify the right function spaces: direct and inverse approximation theorem (Bernstein and Jackson type theorems): f can be approximated by trig polynomials in L^2 to order s iff $f \in H^s, \|f\|_{H^s}^2 = \sum_{k=0}^s \left\|\nabla^k f\right\|_{L^2}^2$. Functions of interest are in the right spaces (PDE theory, real analysis, etc).

Optimal error estimates

- A priori estimates (for piecewise linear finite elements, $\alpha = 1/d, s = 2$)

$$\|f_m - f^*\|_{H^1} \leq Cm^{-\alpha} \|f^*\|_{H^s}$$

- A posteriori estimates (say in finite elements):

$$\|f_m - f^*\|_{H^1} \leq Cm^{-\alpha} \|f_m\|_h$$

Convergence rate is $O(n^{-1/2})$ independent of dimension.

Analyzing the error The empirical risk

$$\mathcal{R}_n(\theta) = \frac{1}{n} \sum_i \left(f\left(\boldsymbol{x}_i, \theta\right) - f^*\left(\boldsymbol{x}_i\right)\right)^2$$

To estimate the population risk:

$$\mathcal{R}(\hat{\theta}) = \mathbb{E}\left(f(\boldsymbol{x}, \hat{\theta}) - f^*(\boldsymbol{x})\right)^2, \quad \hat{\theta} = \operatorname{argmin} \mathcal{R}_n(\theta)$$

(Yaim Cooper: Such $\hat{\theta}$'s form a $m-n$ dimensional (smooth) manifold).

Difficulty: $\hat{\theta}$ is highly correlated with the data set S Study the worse case situation in the hypothesis space. Expect the optimal error rate to be $O(1/m) + O(1/\sqrt{n})$, both are Monte Carlo rates.

Generalization Gap

$\mathcal{R}(\hat{\theta}) - \hat{\mathcal{R}}_n(\hat{\theta}) = I(g) - I_n(g), \quad g(\boldsymbol{x}) = (f(\boldsymbol{x}, \theta) - f^*(\boldsymbol{x}))^2$

$$I(g) = \int_{\mathbb{R}^d} g(\boldsymbol{x}) d\mu, \quad I_n(g) = \frac{1}{n} \sum_j g\left(\boldsymbol{x}_j\right)$$

- For fixed $g = h$, we have

$$|I(h) - I_n(h)| \sim \frac{1}{\sqrt{n}}$$

- For Lipschitz functions (Wasserstein distance)

$$\sup_{\|h\|_{Lip} \leq 1} |I(h) - I_n(h)| \sim \frac{1}{n^{1/d}}$$

- For functions in Barron space, to be defined later

$$\sup_{\|h\|_{\mathcal{B}} \leq 1} |I(h) - I_n(h)| \sim \frac{1}{\sqrt{n}}$$

4 Rademacher Complexity

Let $\mathcal{H}$ be a set of functions, and $S = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ be a set of data points. Then, the Rademacher complexity of $\mathcal{H}$ with respect to S is defined as

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

where $\{\xi_i\}_{i=1}^n$ i.i.d. random variables taking values ± 1 with equal probability.

Theorem 1. (Rademacher complexity and the generalization gap) Given a function class $\mathcal{H}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random samples $S = (x_1, \cdots, x_n)$

$$\begin{array}{l} \sup_{h\in\mathcal{H}} \left| \mathbb{E}_x[h(\boldsymbol{x})] - \frac{1}{n}\sum_{i=1}^n h\left(\boldsymbol{x}_i\right) \right| \le 2\hat{R}_S(\mathcal{H}) + \sup_{h\in\mathcal{H}} \|h\|_\infty \sqrt{\frac{\log(2/\delta)}{2n}} \\ \sup_{h\in\mathcal{H}} \left| \mathbb{E}_x[h(\boldsymbol{x})] - \frac{1}{n}\sum_{i=1}^n h\left(\boldsymbol{x}_i\right) \right| \ge \frac{1}{2}\hat{R}_S(\mathcal{H}) - \sup_{h\in\mathcal{H}} \|h\|_\infty \sqrt{\frac{\log(2/\delta)}{2n}} \end{array}$$

We want large space but low complexity. (Donsker space).
If $\mathcal{H}$ contains a single function, then $\hat{R}_S(\mathcal{H}) \sim O(1/\sqrt{n})$
If $\mathcal{H}$ contains functions that can fit any random values on S, then $\hat{R}_S(\mathcal{H}) \sim O(1)$
If $\mathcal{H}$ = unit ball in Barron space: $\hat{R}_S(\mathcal{H}) \sim O(1/\sqrt{n})$
If $\mathcal{H}$ = unit ball in Lipschitz space: $\hat{R}_S(\mathcal{H}) \sim O\left(1/n^{1/d}\right)$
If $\mathcal{H}$ = unit ball in $C^0 : R_S(\mathcal{H}) \sim O(1)$.
In summary, in low dimension, different approximation methods differ by their order of convergence. In high dimension, different machine learning models differ by the size of the function spaces they are associated with.
Within the right space, we have Monte Carlo convergence rates. Outside the space, we encounter curse of dimensionality.

Identify the right spaces (direct/inverse approx. theorems, complexity control)
- kernel method $\rightarrow$ RKHS space
- shallow networks $\rightarrow$Barron space
- deep residual networks $\rightarrow$ compositional function space

Well-posed models:
- explicit regularization gives rise to optimal error rates
- current results on implicit regularization severely limits the size of the function spaces

A priori estimates for regularized kernel method: fix any $\lambda > 0$,

$$L(f) = \frac{1}{n}\sum_{i=1}^{n} \left(f\left(\boldsymbol{x}_i\right) - f^*\left(\boldsymbol{x}_i\right)\right)^2 + \frac{\lambda}{\sqrt{n}}\|f\|_{\mathcal{H}_k}$$

Theorem 2. Assume $f^* \in \mathcal{H}_k$. Let

$$\hat{f}_n \overset{\text{def}}{=} \operatorname{argmin}_{f\in\mathcal{H}_k} L(f)$$

We have that

$$\mathbb{E}_x \left[\left| \hat{f}_n(\boldsymbol{x}) - f^*(\boldsymbol{x}) \right|^2 \right] \le C(\lambda) \frac{\|f^*\|_{\mathcal{H}_k}}{\sqrt{n}}$$

5 Barron spaces

Two-layer neural networks:

$$\frac{1}{m}\sum_{j=1}^{m} a_j \sigma\left(\boldsymbol{b}_j^T \boldsymbol{x} + c_j\right)$$

Consider the function $f : D_0 = [-1.1]^d \mapsto \mathbb{R}$ of the following form

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

$\Omega = \mathbb{R}^1 \times \mathbb{R}^d \times \mathbb{R}^1, \rho$ is a probability distribution on Ω. Fourier analog: $\rho(da, d\omega) = \delta(a - f(\omega))dad\omega$ (not normalizable).

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} f(\omega)\cos\left(\omega^T \boldsymbol{x}\right) d\omega = \int_{\mathbb{R}^1\times\mathbb{R}^d} a\cos\left(\omega^T \boldsymbol{x}\right)\rho(da, d\omega)$$
$$\|f\|_{\mathcal{B}_p} = \inf_\rho \left(\mathbb{E}_\rho\left[|a|^p \left(\|\boldsymbol{b}\|_1 + |c|\right)^p\right]\right)^{1/p}$$
$$\mathcal{B}_p = \text{ completion of } \left\{f \in \mathcal{S} : \|f\|_{\mathcal{B}_p} < \infty\right\}, \quad \mathcal{B}_\infty \subset \cdots \mathcal{B}_2 \subset \mathcal{B}_1$$

It is a good idea because approximation by two layer networks becomes Monte Carlo integration:

$$f(\boldsymbol{x}) \sim f_m(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma\left(\boldsymbol{b}_j^T \boldsymbol{x} + c_j\right)$$

Theorem 3. (Direct Approximation Theorem) There exists an absolute constant C_0 such that

$$\|f - f_m\|_{L^2(D_0)} \leq \frac{C_0\|f\|_{\mathcal{B}_2}}{\sqrt{m}}$$

Theorem 4. (lnverse Approximation Theorem) For $p > 1$, let

$$\mathcal{N}_{p,C} \stackrel{\text{def}}{=} \left\{\frac{1}{m}\sum_{k=1}^{m} a_k \sigma\left(b_k^T \boldsymbol{x} + c_k\right) : \left(\frac{1}{m}\sum_{k=1}^{m} |a_k|^p \left(\|b_k\|_1 + c_k\right)^p\right)^{1/p} \leq C, m \in \mathbb{N}^+\right\}$$

Let f^* be a continuous function. Assume there exists a constant C and a sequence of functions $f_m \in \mathcal{N}_{p,C}$ such that

$$f_m(\boldsymbol{x}) \to f^*(\boldsymbol{x})$$

for all $x \in D_0$. Then there exists a probability distribution ρ on Ω, such that

$$f^*(\boldsymbol{x}) = \int a\sigma\left(\boldsymbol{b}^T \boldsymbol{x} + c\right)\rho(da, d\boldsymbol{b}, dc)$$

for all $x \in D_0$.

Theorem 5. Let $\mathcal{F}_Q = \{f \in \mathcal{B}_1, \|f\|_{\mathcal{B}_1} \leq Q\}$. Then we have

$$\hat{\mathcal{R}}_n\left(\mathcal{F}_Q\right) \leq 2Q\sqrt{\frac{2\ln(2d)}{n}}$$

...

6 Compositional Function Space

$$\boldsymbol{z}_{l+1,L}(\boldsymbol{x}) = z_{l,L}(\boldsymbol{x}) + \frac{1}{L}\boldsymbol{U}_l\sigma \circ (\boldsymbol{W}_{l^2l,L}(\boldsymbol{x}))$$
$$\boldsymbol{z}_{0,L}(\boldsymbol{x}) = \boldsymbol{V}\boldsymbol{x}, \quad f(\boldsymbol{x},\theta) = \alpha \cdot \boldsymbol{z}_{L,L}(\boldsymbol{x})$$

A discrete flow with RHS in Barron space. Let $\{\rho_t\}$ be a family of prob distributions (for $(\boldsymbol{U},\boldsymbol{W})$) such that $\mathbb{E}_{\rho_t}, g(\boldsymbol{U},\boldsymbol{W})$ is integrable as a function of t for any continuous function g. Define:

$$\boldsymbol{z}(\boldsymbol{x},0) = \boldsymbol{V}\boldsymbol{x}$$
$$\frac{d}{dt}\boldsymbol{z}(\boldsymbol{x},t) = \mathbb{E}_{(\boldsymbol{U},\boldsymbol{W})\sim\rho_t}\boldsymbol{U}\sigma \circ (\boldsymbol{W}\boldsymbol{z}(\boldsymbol{x},t))$$

Let $f_{\alpha,\{\rho_t\},V}(\boldsymbol{x}) = \alpha^T\boldsymbol{z}(\boldsymbol{x},1)$ and define

$$\frac{d}{dt}\mathbf{N}_p(t) = (\mathbb{E}_{\rho t}|\boldsymbol{U}|^p|\boldsymbol{W}|^p)^{1/p}\,\mathbf{N}_p(t), \quad \mathbf{N}_p(0) = \mathbf{I}$$
$$\|f\|_{\mathcal{D}_p} = \inf_{f=f_{\alpha,\{\rho_t\},V}} \|\alpha\|_p \left\|\mathbf{N}_p(1)\right\|_{p,p} \|\boldsymbol{V}\|_{p,p}$$

The matrix operations are done entry-wise.

$$f(\boldsymbol{x}) = \textstyle\int_\Omega a\sigma\left(\boldsymbol{b}^T\boldsymbol{x}+c\right)\rho(da,d\boldsymbol{b},dc)$$
$$\tfrac{d}{dt}z(\boldsymbol{x},t) = \mathbb{E}_{\rho(a,b,c)}\begin{bmatrix} a \\ 0 \\ 0 \end{bmatrix}\sigma\circ\left(\left[0,b^T,c\right]z(\boldsymbol{x},t)\right),$$
$$z(\boldsymbol{x},0) = \begin{bmatrix} 0 \\ \boldsymbol{x} \\ 1 \end{bmatrix}$$

Then, $f(\boldsymbol{x}) = e_1^T z(\boldsymbol{x},1)$. $\mathcal{B}_2 \subset \mathcal{D}_2$. There exists constant $C>0$, such that

$$\|f\|_{\mathcal{D}_2} \le \sqrt{d+1}\|f\|_{\mathcal{B}_2}$$

holds for any $f \in \mathcal{B}_2$...

7 Regularized Model and a Priori Estimates

Regularized loss function:

$$J(\theta) = L(\theta) + \lambda\left(\|\theta\|_{\mathcal{D}_1}+1\right)\sqrt{\tfrac{2\log(2d)}{n}}$$
$$\|\theta\|_P = \left\||\alpha|^T\left(I+\tfrac{1}{L}\left|\boldsymbol{U}_{L-1}\|\boldsymbol{W}_{L-1}\right|\right)\cdots\left(I+\tfrac{1}{L}\left|\boldsymbol{U}_1\|\boldsymbol{W}_1\right|\right)|\boldsymbol{V}|\right\|_1$$

Theorem 6. (A-priori estimate) Assume that $f^\star : [-1,1]^d \to [-1,1]$ such that $f^* \in \mathcal{D}_2$. Let

$$\hat{\theta} = \mathrm{argmin}_\theta\, J(\theta)$$

Then there exist fixed constants C_0 and L_0 such that if $\lambda > C_0$ and $L > L_0$, then for any $\delta > 0$, with probability at least $1-\delta$,

$$L(\hat{\theta}) \lesssim \frac{\|f^*\|_{\mathcal{D}_2}^2}{L} + \lambda\left(\|f^*\|_{\mathcal{D}_1}^3+1\right)\sqrt{\frac{\log(2d)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}$$

Gradient descent for unregularized model Questions: can gradient descent algorithm achieve global minima for the empirical risk? can the solution obtained generalize (small population risk)? "implicit regularization". Over-parametrized regime $m \geq O\left(n^{\alpha}\right)$: CLT scaling

$$f_0(\boldsymbol{x}) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k^0 \sigma\left(\boldsymbol{x}^T \boldsymbol{b}_k^0\right)$$

Optimization: Empirical risk goes to 0?
With random (normal) initialization, GD can fit any target (random labels).
Generalization: Population risk (real error) small?
Error estimates involve ε or κ. ...