Generalization error analysis of deep learning via a kernel perspective

Department of Mathematical Informatics, The University of Tokyo. Taiji Suzuki.

We develop a new theoretical framework to analyze the generalization error of deep learning, and derive a new fast learning rate for two representative algorithms: *empirical risk minimization* and *Bayesian deep learning*. The series of theoretical analyses of deep learning has revealed its high expressive power and universal approximation capability. Although these analyses are highly nonparametric, existing generalization error analyses have been developed mainly in a fixed dimensional parametric model. To compensate this gap, we develop an infinite dimensional model that is based on an integral form as performed in the analysis of the universal approximation capability. This allows us to define a reproducing kernel Hilbert space corresponding to each layer. Our point of view is to deal with the ordinary finite dimensional deep neural network as a finite approximation of the infinite dimensional one. The approximation error is evaluated by the *degree of freedom* of the reproducing kernel Hilbert space in each layer. To estimate a good finite dimensional model, we consider both of empirical risk minimization and Bayesian deep learning. We derive its generalization error bound and it is shown that there appears bias-variance trade-off in terms of the number of parameters of the finite dimensional approximation. We show that the optimal width of the internal layers can be determined through the degree of freedom and the convergence rate can be faster than $O(1/\sqrt{n})$ rate.

We define a feature space on the ℓ-th layer. The feature space is a a probability space $(\mathcal{T}_\ell, \mathcal{B}_\ell, \mathcal{Q}_\ell)$ where $\mathcal{T}_\ell$ is a Polish space, $\mathcal{B}_\ell$ is its Borel algebra, and $\mathcal{Q}_\ell$ is a probability measure on $(\mathcal{T}_\ell, \mathcal{B}_\ell)$. Now the input x is a d_{x}-dimensional real vector, and thus we may set $\mathcal{T}_1 = \{1, \dots, d_{\mathrm{x}}\}$. Since the output is one dimensional, the output layer is just a singleton $\mathcal{T}_{L+1} = \{1\}$. Based on these feature spaces, our integral form of the deep neural network is constructed by stacking the map on the ℓ-th layer $f_\ell^{\mathrm{o}} : L_2(Q_\ell) \to L_2(Q_{\ell+1})$ given as

$$f_\ell^{\mathrm{o}}[g](\tau) = \int_{\mathcal{T}_\ell} h_\ell^{\mathrm{o}}(\tau, w)\eta(g(w))\mathrm{d}Q_\ell(w) + b_\ell^{\mathrm{o}}(\tau),$$

where η is an activation function, $h_\ell^{\mathrm{o}}(\tau, w)$ corresponds to the weight of the feature w for the output τ and $h_\ell^{\mathrm{o}} \in L_2(Q_{\ell+1} \times Q_\ell)$ and $h_\ell^{\mathrm{o}}(\tau, \cdot) \in L_2(Q_\ell)$ for all $\tau \in \mathcal{T}_{\ell+1}$. Specifically, the first and the last layers are represented as $f_1^{\mathrm{o}}[x](\tau) = \sum_{j=1}^{d_{\mathrm{x}}} h_1^{\mathrm{o}}(\tau, j)x_j Q_1(j) + b_1^{\mathrm{o}}(\tau)$, and $f_L^{\mathrm{o}}[g](1) = \int_{\mathcal{T}_L} h_L^{\mathrm{o}}(w)\eta(g(w))\mathrm{d}Q_L(w) + b_L^{\mathrm{o}}$ where we wrote $h_L^{\mathrm{o}}(w)$ to indicate $h_L^{\mathrm{o}}(1, w)$ for simplicity. Then the true function f^{o} is given as $f^{\mathrm{o}}(x) = f_L^{\mathrm{o}} \circ f_{L-1}^{\mathrm{o}} \circ \cdots \circ f_1^{\mathrm{o}}(x)$. We want to approximate this infinite dimensional model by a finite dimensional one which is defined by using $W^{(\ell)} \in \mathbb{R}^{m_{\ell+1} \times m_\ell}$ as

$$f_\ell^*(g) = W^{(\ell)}\eta(g) + b^{(\ell)} \quad (g \in \mathbb{R}^{m_\ell},\ \ell = 1, \dots, L), \qquad f^*(x) = f_L^* \circ f_{L-1}^* \circ \cdots \circ f_1^*(x).$$

Let the output of the ℓ-th layer be $F_\ell^{\mathrm{o}}(x, \tau) := (f_\ell^{\mathrm{o}} \circ \cdots \circ f_1^{\mathrm{o}}(x))(\tau)$. We define a *reproducing kernel Hilbert space* corresponding to the ℓ-th layer ($\ell \geq 2$) by introducing its associated kernel function $\mathsf{k}_\ell : \mathbb{R}^{d_{\mathrm{x}}} \times \mathbb{R}^{d_{\mathrm{x}}} \to \mathbb{R}$ as

$$\mathsf{k}_\ell(x, x') := \int_{\mathcal{T}_\ell} \eta(F_{\ell-1}^{\mathrm{o}}(x, \tau))\eta(F_{\ell-1}^{\mathrm{o}}(x', \tau))\mathrm{d}Q_\ell(\tau).$$

Let the *degree of freedom* be $N_\ell(\lambda) = \sum_{j=1}^{\infty} \frac{\mu_j^{(\ell)}}{\mu_j^{(\ell)} + \lambda}$ for $\lambda > 0$ where $\mu_1^{(\ell)} \geq \mu_2^{(\ell)} \geq \dots$ be the eigenvalues of the kernel in $L_2(P_X)$.

Theorem. Under suitable conditions such as $\|h_\ell^{\mathrm{o}}(\tau, \cdot)\|_{L_2(Q_\ell)} \leq R$ $(\forall \tau, \ell)$, there exist constants $C_1, C_2 > 0$ such that, if $m_\ell \geq C_1 N_\ell(\lambda_\ell)\log\left(N_\ell(\lambda_\ell)\right)$ $(\ell = 2, \dots, L)$, then, for any $\lambda_\ell > 0$ $(\ell = 2, \dots, L)$, it holds that

$$\|\widehat{f} - f^{\mathrm{o}}\|_{L_2(P_X)}^2 \leq C_2 \left[\left(\sum_{\ell=2}^{L} \sqrt{\lambda_\ell}\right)^2 + \frac{1}{n}\sum_{\ell=1}^{L} m_\ell m_{\ell+1}\log(n)^2\right]$$

with high probability. By balancing the first and second terms, we obtain a fast learning rate.