カーネル法の理論による深層学習の汎化誤差解析
(Generalization error analysis of deep learning via kernel method theory)

Taiji Suzuki TAIJI@MIST.I.U-TOKYO.AC.JP
Department of Mathematical Informatics
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan,
PRESTO, Japan Science and Technology Agency,
Center for Advanced Integrated Intelligence Research, RIKEN

Abstract

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 [1].

1. Introduction

Deep learning has been showing great success in several applications such as computer vision, natural language processing, and many other area related to pattern recognition. Several high-performance methods have been developed and it has been revealed that deep learning possesses great potential. Despite the development of practical methodologies, its theoretical understanding is not satisfactory. Wide rage of researchers including theoreticians and practitioners are expecting deeper understanding of deep learning.

Among theories of deep learning, a well developed topic is its expressive power. It has been theoretically shown that deep neural network has exponentially large expressive power against the number of layers. For example, Montufar et al. (2014) showed that the number of polyhedral regions created by deep neural network can exponentially grow as the number of layers increases. Bianchini and Scarselli (2014) showed that the Betti numbers of the level set of a function created by deep neural network grows up exponentially against the number of layers. Other researches also concluded similar facts using different notions such as tensor rank and extrinsic curvature (Cohen et al., 2016; Cohen and Shashua, 2016; Poole et al., 2016).

Another important issue in neural network theories is its universal approximation capability. It is well known that 3-layer neural networks have the ability, and thus the deep neural network also

[1] The extended version of this article can be found in Suzuki (2017).

does (Cybenko, 1989; Hornik, 1991; Sonoda and Murata, 2015). When we discuss the universal approximation capability, the target function that is approximated is arbitrary and the theory is highly nonparametric in its nature.

Once we knew the expressive power and universal approximation capability of deep neural network, the next theoretical question naturally arises in its generalization error. The generalization ability is typically analyzed by evaluating the *Rademacher complexity*. Bartlett (1998) studied 3-layer neural networks and characterized its Rademacher complexity using the norm of weights. Koltchinskii and Panchenko (2002) studied deep neural network and derived its Rademacher complexity under norm constraints. More recently, Neyshabur et al. (2015) analyzed the Rademacher complexity based on more generalized norm, and Sun et al. (2015) derived a generalization error bound with a large margine assumption. As a whole, the studies listed above derived $O(1/\sqrt{n})$ convergence of the generalization error where n is the sample size. Although this is minimax optimal, it is expected that we could show faster convergence rate with some additional assumptions such as strong convexity of the loss function. Moreover, the generalization error bound has been mainly given in finite dimensional models. As we have observed, the deep neural network possesses exponential expressive power and universal approximation capability which are highly nonparametric characterizations. This means that the theories are developed separately in the two regimes; finite dimensional parametric model and infinite dimensional nonparametric model. Therefore, theories that connect these two regimes are required to comprehensively understand statistical performance of deep learning.

In this study, we consider both of empirical risk minimization and Bayesian deep learning and analyze the generalization error using the terminology of kernel methods. Consequently, (i) we derive a faster learning rate than $O(1/\sqrt{n})$ and (ii) we connect the finite dimensional regime and the infinite dimensional regime based on the theories of kernel methods. To analyze a sharper generalization error bound, we utilize the so-called local Rademacher complexity technique for the empirical risk minimization method (Mendelson, 2002; Bartlett et al., 2005; Koltchinskii, 2006; Giné and Koltchinskii, 2006), and, as for the Bayesian method, we employ the theoretical techniques developed to analyze nonparametric Bayes methods (Ghosal et al., 2000; van der Vaart and van Zanten, 2008, 2011). The obtained generalization error bound is summarized in Table 1[2].

2. Generalization error analysis of deep learning

We consider a regression model formulated as

$$y_i = f^{\mathrm{o}}(x_i) + \xi_i \ \ (i = 1, \dots, n),$$

where $(\xi_i)_{i=1}^n$ is an i.i.d. sequence of Gaussian noises $N(0, \sigma^2)$ with mean 0 and variance σ^2, and $(x_i)_{i=1}^n$ is generated independently identically from a distribution $P(X)$ with a compact support in $\mathbb{R}^{d_{\mathrm{x}}}$. We suppose that f^{o} has a hierarchic structure which is defined below. 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

[2] $a \vee b$ indicates $\max\{a, b\}$.

Table 1: Summary of derived bounds for the generalization error $\|\widehat{f} - f^{\mathrm{o}}\|_{L_2(P(X))}^2$ where n is the sample size, R is the norm of the weight in the internal layers, $\hat{R}_\infty$ is an L_∞-norm bound of the functions in the model, σ is the observation noise, d_{x} is the dimension of the input, m_ℓ is the width of the ℓ-th internal layer and $N_\ell(\lambda_\ell)$ for ($\lambda_\ell > 0$) is the degree of freedom (Eq. (1)). More details can be found in Suzuki (2017).

	Error bound
General setting	$L\sum_{\ell=2}^{L} R^{L-\ell+1}\lambda_\ell + \frac{\sigma^2+\hat{R}_\infty^2}{n}\sum_{\ell=1}^{L} m_\ell m_{\ell+1}\log(n)$ under an assumption that $m_\ell \gtrsim N_\ell(\lambda_\ell)\log(N_\ell(\lambda_\ell))$.
Finite dimensional model	$\frac{\sigma^2+\hat{R}_\infty^2}{n}\sum_{\ell=1}^{L} m_\ell^* m_{\ell+1}^*\log(n)$ where m_ℓ^* is the true width of the ℓ-th internal layer.
Polynomial decay eigenvalue	$L\sum_{\ell=2}^{L}(R\vee 1)^{L-\ell+1} n^{-\frac{1}{1+2s_\ell}}\log(n) + \frac{d_{\mathrm{x}}^2}{n}\log(n)$ where s_ℓ is the decay rate of the eigenvalue of the kernel function on the ℓ-th layer.

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

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

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

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* (RKHS) 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

$$N_\ell(\lambda) = \textstyle\sum_{j=1}^{\infty} \mu_j^{(\ell)}/(\mu_j^{(\ell)} + \lambda) \tag{1}$$

for $\lambda > 0$ where $\mu_1^{(\ell)} \geq \mu_2^{(\ell)} \geq \dots$ be the eigenvalues of the kernel in $L_2(P_X)$.

We assume that there exist R and R_b such that the true function satisfies the following condition: $\|h_\ell^{\mathrm{o}}(\tau, \cdot)\|_{L_2(Q_\ell)} \leq R\ (\forall \tau \in T_\ell)$, $|b_\ell^{\mathrm{o}}(\tau)| \leq R_b\ \ (\forall \tau \in T_\ell)$. Let $\bar{R} = 2R$ and $\bar{R}_b = 2R_b$. Under this condition, we construct an estimator in the following finite dimensional model:

$$\begin{aligned}\mathcal{F} =& \{f(x) = (W^{(L)}\eta(\cdot) + b^{(L)}) \circ \cdots \circ (W^{(1)}x + b^{(1)}) \mid W^{(\ell)} \in \mathbb{R}^{m_{\ell+1} \times m_\ell},\ b^{(\ell)} \in \mathbb{R}^{m_{\ell+1}}, \\ & \|W^{(\ell)}\|_{\mathrm{F}} \leq \bar{R},\ \|b^{(\ell)}\| \leq \bar{R}_b\ (\ell = 1, \dots, L)\}.\end{aligned}$$

The empirical risk minimizer is given as $\widehat{f} := \mathrm{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n}(y_i - f(x_i))^2$. We can also construct a Bayes estimator by putting a prior distribution Π on $\mathcal{F}$. Then we obtain an upper-bound of the generalization errors of the empirical risk minimizer and the Bayes estimator as in the following theorem.

Theorem 1 (Informal) *Let $\widehat{f}$ be either of the empirical risk minimizer and the Bayes estimator with an appropriate prior. Then, under some technical conditions, 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, with high probability,*

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

References

P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33: 1487–1537, 2005.

P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

M. Bianchini and F. Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25(8):1553–1565, 2014.

N. Cohen and A. Shashua. Convolutional rectifier networks as generalized tensor decompositions. In *Proceedings of the 33th International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 955–963, 2016.

N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Proceedings of the 29th Annual Conference on Learning Theory*, pages 698–728, 2016.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems 27*, pages 2924–2932. 2014.

B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1376–1401, 2015.

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*, pages 3360–3368. 2016.

S. Sonoda and N. Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 2015.

S. Sun, W. Chen, L. Wang, and T.-Y. Liu. Large margin deep neural networks: theory and algorithms. *arXiv preprint arXiv:1506.05232*, 2015.

T. Suzuki. Fast learning rate of deep learning via a kernel perspective. *arXiv preprint arXiv:1705.10182*, 2017.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.