An Information-Theoretic View for Deep Learning

Jingwei Zhang[*]   Tongliang Liu[*]   Dacheng Tao[*]

Abstract

Deep learning has transformed the computer vision, natural language processing and speech recognition. However, the following two critical questions are remaining obscure: (1) why deep neural networks generalize better than shallow networks? (2) Does it always hold that a deeper network leads to better performance? Specifically, letting L be the number of convolutional and pooling layers in a deep neural network, and n be the size of the training sample, we derive the upper bound on the expected generalization error for this network, i.e.,

$$\mathbb{E}[R(W) - R_S(W)] \leq \exp\left(-\frac{L}{2}\log\frac{1}{\eta}\right)\sqrt{\frac{2\sigma^2}{n}I(S, W)}$$

where $\sigma > 0$ is a constant depending on the loss function, $0 < \eta < 1$ is a constant depending on the information loss for each convolutional or pooling layer, and $I(S, W)$ is the mutual information between the training sample S and the output hypothesis W. This upper bound discovers: (1) As the network increases its number of convolutional and pooling layers L, the expected generalization error will decrease exponentially to zero. Layers with strict information loss, such as the convolutional layers, reduce the generalization error of deep learning algorithms. This answers the first question. However, (2) algorithms with zero expected generalization error does not imply a small test error or $\mathbb{E}[R(W)]$. This is because $\mathbb{E}[R_S(W)]$ will be large when the information for fitting the data is lost as the number of layers increases. This suggests that the claim "the deeper the better" is conditioned on a small training error or $\mathbb{E}[R_S(W)]$.

---

[*]UBTECH Sydney AI Centre and the School of Information Technologies in the Faculty of Engineering and Information Technologies at The University of Sydney, NSW, 2006, Australia, zjin8228@uni.sydney.edu.au, tongliang.liu@sydney.edu.au, dacheng.tao@sydney.edu.au

arXiv:1804.09060v1 [stat.ML] 24 Apr 2018

1 Introduction

We study the standard framework of statistical learning where the instance space is denoted by $\mathcal{Z}$ and the hypothesis space is denoted by $\mathcal{W}$. We denote the training sample by an n-tuple $S = (Z_1, Z_2, ..., Z_n)$ where each element Z_i is drawn i.i.d. from an unknown distribution D. A learning algorithm $\mathcal{A} : S \to \mathcal{W}$ can be seen as a randomized mapping from the training sample space $\mathcal{Z}^n$ to the hypothesis space $\mathcal{W}$. We characterize the learning algorithm $\mathcal{A}$ by a Markov kernel $P_{W|S}$, which means that given the training sample S, the algorithm picks a hypothesis in $\mathcal{W}$ according to the conditional distribution $P_{W|S}$.

We introduce a loss function $\ell : \mathcal{W} \times Z \mapsto \mathbb{R}^+$ to measure the quality of a prediction w.r.t. a hypothesis. For any learned hypothesis W by S, we define the expected risk

$$R(W) = \mathbb{E}_{Z \sim D}[\ell(W, Z)] \, , \tag{1}$$

and the empirical risk

$$R_S(W) = \frac{1}{n} \sum_{i=1}^{n} \ell(W, Z_i) \, . \tag{2}$$

For a learning algorithm $\mathcal{A}$, the generalization error is defined as

$$G_S(D, P_{W|S}) = R(W) - R_S(W) \, . \tag{3}$$

A small generalization error implies the learned hypothesis will have similar performances on the training and test datasets.

In this paper, we study the following expected generalization error for deep learning algorithms

$$G(D, P_{W|S}) = \mathbb{E}[R(W) - R_S(W)] \, , \tag{4}$$

where the expectation is over the joint distribution $P_{W,S} = D^n \times P_{W|S}$.

We have the following decomposition

$$\mathbb{E}[R(W)] = G(D, P_{W|S}) + \mathbb{E}[R_S(W)] \tag{5}$$

where the first term on the right hand side is the expected generalization error and the second term reflects how well the learned hypothesis fit the training data from an expectation view.

When designing a learning algorithm, we want the expectation of the expected risk, i.e., $\mathbb{E}[R(W)]$, to be as small as possible. However, it is not easy to have small values for the expected generalization error $G(D, P_{W|S})$ and the expected empirical risk $\mathbb{E}[R_S(W)]$ at the same time. Usually, if a model fits the training data too well, it may generalize poorly on the test data. This is known as the bias-variance trade-off problem (Domingos 2000). Surprisingly, deep learning algorithms have successfully and empirically shown their power in minimizing $G(D, P_{W|S})$ and $\mathbb{E}[R_S(W)]$ at the same time. They have small $\mathbb{E}[R_S(W)]$ because neural networks having deep architectures are efficient to compactly

represent highly-varying functions (Sonoda and Murata 2015). However, the theoretical justification for their small expected generalization errors $G(D, P_{W|S})$ remains elusive.

In this paper, we study the expected generalization error for deep learning algorithms from an information-theoretic point of view. We will show that as the number of layers grows, the expected generalization error $G(D, P_{W|S})$ will decrease exponentially to zero. Specifically, in Theorem 2, we prove that

$$G(D, P_{W|S}) = \mathbb{E}[R(W) - R_S(W)] \leq \exp\left(-\frac{L}{2}\log\frac{1}{\eta}\right)\sqrt{\frac{2\sigma^2}{n}I(S, W)},$$

where L is the number of information loss layers of deep neural networks, $0 < \eta < 1$ is a constant depending on the information loss of each layer, $\sigma > 0$ is a constant depending on the loss function, n is the size of the training sample S, and $I(S, W)$ is the mutual information between the input training sample S and the output hypothesis W.

Our conclusion is based on two important results related to information theory. The first is the strong data processing inequalities (SDPIs) proposed by (Ahlswede and Gács 1976), which states that for a Markov chain $U \to V \to W$, if there is information loss on the mapping $V \to W$, then $I(U, V) \leq \eta I(U, W)$ and $\eta < 1$ is a nonnegative information loss factor. Another result is by the line of (Russo and Zou 2015, Xu and Raginsky 2017), which states that the mutual information between the input and output of a learning algorithm controls its generalization error.

Our result is not conflicted with the bias-variance trade-off. Although the expected generalization error will decrease exponentially to zero as the number of information loss layers increases, the expected empirical risk $\mathbb{E}[R_S(W)]$ will increase as the information loss will be harmful to the fitting of the training data. This implies that, when designing deep learning algorithms, more efforts should be placed on the balance of the information loss and the training error.

The advantage of using the mutual information between the input and output to bound the expected generalization error is that it depends on almost every aspects of the learning algorithm, including the data distribution, the complexity of hypothesis class, and the property of the learning algorithm itself; While the traditional frameworks for proving PAC-learnablility (Mohri et al. 2012) may only focus some of the aspects. For example, the VC dimension (Vapnik 2013), covering number (Zhang 2002), Rademacher (Bartlett and Mendelson 2002, Bartlett et al. 2005, Liu et al. 2017), PAC-Bayes (Langford and Shawe-Taylor 2003), algorithmic stability (Liu et al. 2017, Bousquet and Elisseeff 2002) and robustness (Xu and Mannor 2012) based frameworks.

The rest of this paper is organized as follows: in Section 2, we relate our DNNs with a Markov chain; Section 3 will exploit the strong data processing inequality to derive how the mutual information between intermediate features representations and the output varies in deep neural networks; our main results will be given in Section 4, which gives the exponential generalization error bound for DNNs in terms of depth L; then we give the proofs of our main theorem in Section 5; finally, we conclude our paper and point out some implications in Section 6 .
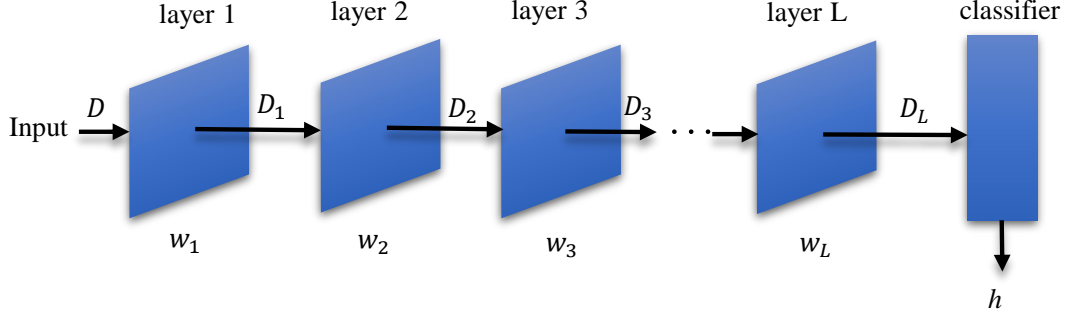
Figure 1: Hierarchical Feature Mapping of Deep Neural Networks with L Hidden Layers

$$S \to T_1 \to T_2 \to T_3 \to \cdots \to T_L$$
$$\downarrow$$
$$W \to [w_2, \ldots, w_L; h] \to [w_3, \ldots, w_L; h] \to \cdots \to [w_L; h] \to h$$

Figure 2: The Feature Mapping of Deep Neural Networks Forms a Markov Chain

2 The Hierarchical Feature Mapping of DNNs and Its Relations to Markov Chain

Let's introduce some notations for deep neural networks (DNNs). As shown in figure 1, a DNN with L hidden layers can be seen as L feature maps that conduct feature transformations L times on the input Z sequentially. After feature transformations L times, the learned feature will be the input of a classifier (or regressor) at the output layer. If the distribution on a single input is D, then we denote the distribution after going through k-th hidden layer as D_k and the corresponding variable is Z_k where $k = 1, \ldots, L$ and the weight of the whole network is denoted by $W = [w_1, \ldots, w_L; h] \in \mathcal{W}$, where $\mathcal{W}$ is the space of all possible weights. As shown in figure 2, the input S is transformed layer by layer and the output of the k-th hidden layer is T_k, where $k = 1, \ldots, L$. We also denote the j-th sample after going through the k-th hidden layer by Z_{k_j}. In words, we have the following relationship:

$$Z \sim D \tag{6}$$
$$Z_k \sim D_k, \quad for \quad k = 1, \ldots, L \tag{7}$$
$$S = \{Z_1, \ldots, Z_n\} \sim D^n \tag{8}$$
$$T_k = \{Z_{k_1}, \ldots, Z_{k_n}\} \sim D_k^n, \quad for \quad k = 1, \ldots, L. \tag{9}$$

We now have a Markov model for DNNs, as shown in figure 2. From the Markov property, we know that if $U \to V \to W$ forms a Markov chain, then W is conditionally independent with U given V. Furthermore, from the data processing inequality (Cover and Thomas 2012), we have $I(U, W) \leq I(U, V)$ and the equality holds if and only if $U \to W \to V$ also forms a Markov chain. Applying the data processing inequality to the Markov chain, we have,

$$I(T_L, h) \leq I(T_{L-1}, h) \leq I(T_{L-2}, h) \leq \ldots \leq I(S, h) \leq I(S, W) . \tag{10}$$

This means that the mutual information between input and output is non-increasing as it goes through the network layer by layer. As the feature map in each layer is likely to be non-invertible, the mutual information between the input and output is likely to be strictly decreasing when it goes through each layers. This encourages the study of the strong data processing inequality (Polyanskiy and Wu 2015). In the next section, we will prove that the strong data processing inequality holds for DNNs generally.

3 Information Loss in DNNs

In the previous section, we have modeled the DNN as a Markov chain and, by using the data processing inequality, we conclude that the mutual information between input and output in DNNs is non-increasing. The equalities in equation (10) will not hold for most cases and therefore we can apply the strong data processing inequality to achieve tighter inequalities.

For a Markov chain $U \to V \to W$, the random transformation $P_{V|W}$ can be seen as a channel from an information-theoretic point of view. Strong data processing inequalities (SDPIs) quantify an intuitive observation that the noise inside the channel $P_{V|W}$ will reduce the mutual information between U and W. That is, there exists $0 \leq \eta < 1$, such that

$$I(U, W) \leq \eta I(V, W) . \tag{11}$$

Formally,

Theorem 1 (Ahlswede and Gács 1976). *Consider a Markov chain $W \to X \to Y$ and the corresponding random mapping $P_{Y|X}$. If the mapping $P_{Y|X}$ is not noiseless, that is we cannot recover any X perfectly with probability 1 from the observed random variable $\widehat{Y}$. Then there exists $0 \leq \eta < 1$, such that*

$$I(W, Y) \leq \eta I(X, Y) \tag{12}$$

More details can be found in the comprehensive survey about SDPIs (Polyanskiy and Wu 2015).

Let's consider the k-th hidden layer ($1 \leq k \leq L$) in figure 1. It can be seen as a randomized transformation $P_{Z_k|Z_{k-1}}$ that maps from one distribution D_{k-1} to another

5

distribution D_k (when $k = 1$, we denote $D = D_0$). We then denote the parameters of the k-th hidden layer by w_k. Without loss of generality, let w_k be a matrix of dimension $\mathbb{R}^{d_{k-1} \times d_k}$. Also, we denote the activation function in this layer by $\sigma_k(\cdot)$.

We give the definition of the *contraction layer*.

Definition 1 (Contraction Layer). *A layer in deep neural networks is called a contraction layer if it causes information loss.*

We now give the first result of this paper, which quantifies the information loss in DNNs.

Corollary 1 (Information Loss in DNNs). *Consider a DNN as shown in figure 1 and its corresponding Markov model in figure 2. If its k-th ($1 \leq k \leq L$) hidden layer is a contraction layer, then there exists $0 \leq \eta_k < 1$, such that*

$$I(T_k, h) \leq \eta_k I(T_{k-1}, h) . \tag{13}$$

We show that the most used convolutional or pooling layers are contraction layers.

Lemma 1. *For any layer in a DNN, with parameters $w_k \in \mathbb{R}^{d_{k-1} \times d_k}$, if $rank(w_k) < d_k$, it is a contraction layer.*

Proof. For the k-th hidden layer, consider any input $x_{k-1} \sim D_{k-1}$ and the corresponding output of the k-th hidden layer $x_k \sim P_{D_k|D_{k-1}}$, we have

$$x_k = \sigma_k(w_k x_{k-1}) . \tag{14}$$

As $rank(w_k) < d_k$, then the dimension of its right null space is greater than or equal to 1. We denote the right null space of w_k by $RNULL(w_k)$, then we can pick a vector $\alpha \in RNULL(w_k)$ such that $w_k \alpha = 0$.

Then, we have

$$\sigma_k(w_k(x_{k-1} + \alpha)) = \sigma_k(w_k x_{k-1}) = x_k . \tag{15}$$

Therefore, for any input $x_{k-1} \sim D_{k-1}$ of the k-th layer, there exists $x'_{k-1} = x_{k-1} + \alpha$ such that their corresponding outputs are the same, which means, for any x_{k-1}, we cannot recover it perfectly with probability 1.

We conclude that the mapping $P_{D_k|D_{k-1}}$ is noisy and the corresponding layer will cause information loss. $\square$

Theorem 1 shows that the mutual information $I(T_{k-1}, h)$ will decrease after it goes through a contraction layer. From Lemma 1, we know that the convolutional and pooling layer is often guaranteed to be a contraction layer. For a fully connected layer, the contraction property may not hold as the weight sometimes may be of full rank which leads to a noiseless and invertible mapping. However, the active function employed sub-sequentially can contribute to forming a contraction layer. Without loss of generality, we let all L hidden layers be contraction layers, e.g., convolutional or pooling layers.

4 Exponential Bounds on the Generalization Error of DNNs

Before we introduce our main theorem, we need to restrict the loss function $\ell(W, Z)$ to be σ-sub-Gaussian with respect to Z for any $W \in \mathcal{W}$.

Definition 2 (σ-sub-Gaussian). *A random variable X is said to be σ-sub-Gaussian if the following inequality holds for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp\left(\lambda(X - \mathbb{E}[X])\right)] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \tag{16}$$

We now present our main theorem, which gives an exponential bound for the expected generalization error of deep learning algorithms.

Theorem 2. *For a DNN with L hidden layers, input S, and parameters W. Assuming that the loss function $\ell(W, Z)$ is σ-sub-Gaussian with respect to Z for any $W \in \mathcal{W}$. Without loss of generality, let all L hidden layers are contraction layers (e.g., convolutional or pooling layers). Then, the expected generalization error of the corresponding deep learning algorithm can be upper bounded as follows,*

$$\mathbb{E}[R(W) - R_S(W)] \leq \exp\left(-\frac{L}{2}\log\frac{1}{\eta}\right)\sqrt{\frac{2\sigma^2}{n}I(S, W)} \tag{17}$$

where $\eta < 1$ is the maximum information loss factor for all L contraction layers, that is

$$\eta = \max_{i \in \{1,\ldots,L\}} \eta_i . \tag{18}$$

The upper bound in Theorem 2 is loose w.r.t. the mutual information $I(S, W)$ since we used the inequality $I(S, h) \leq I(S, W)$. However, we have that

$$I(S, W) \leq \min\{H(S), H(W)\} \leq H(S) , \tag{19}$$

which implies that as the number of contraction layers L increases, the expected generalization error will decrease exponentially to zero. We give the detailed proofs of the main theorem in the next section.

From Theorem 2, it implies that deeper neural networks will lead to better generalization error. However, it does not means that the deeper the better. Recall that $\mathbb{E}[R(W)] = G(D, P_{W|S}) + \mathbb{E}[R_S(W)]$. A small $G(D, P_{W|S})$ does not imply a small $\mathbb{E}[R(W)]$. Because the expected empirical training error $\mathbb{E}[R_S(W)]$ will increase due to information loss. Specially, if the information that cares the relationship between the observation X and the target Y loses, fitting the training data will become difficult and the empirical training error will increase. Our results point out a research direction for designing deep learning algorithms that we should increase the number of contraction layers while keep the empirical training errors to be small as well.

The information loss factor η plays an essential role in the generalizing of deep learning algorithms. A successful deep learning algorithm should filter redundant information as much as possible while keep sufficient information to fit the training data. The functions of some deep learning tricks, such as convolution, pooling, and activation, serve very well on filter some redundant information. This further confirms the information-bottle theory (Shwartz-Ziv and Tishby 2017) by that with more contraction layers, more redundant information will be removed while the prediction information is preserved.

5 Proof of Theorem 2

First, by the law of total expectation, we have,

$$\mathbb{E}[R(W) - R_S(W)] = \mathbb{E}\left[\mathbb{E}[R(W) - R_S(W)|w_1, \ldots, w_L]\right] . \tag{20}$$

We now give an upper bound on $\mathbb{E}[R(W) - R_S(W)|w_1, \ldots, w_L]$ similar to the way of (Russo and Zou 2015, Xu and Raginsky 2017).

Lemma 2. *Under the same conditions as in Theorem 2, the upper bound of $\mathbb{E}[R(W) - R_S(W)|w_1, \ldots, w_L]$ is given by*

$$\mathbb{E}[R(W) - R_S(W)|w_1, \ldots, w_L] \leq \sqrt{\frac{2\sigma^2}{n} I\left(T_L, h\right)} . \tag{21}$$

Proof. We have,

$$
\begin{aligned}
&\mathbb{E}[R(W) - R_S(W)|w_1, \ldots, w_L] \\
&= \mathbb{E}_{h,S}\left[\mathbb{E}_{Z\sim D}[\ell(W, Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(W, Z_i)|w_1, \ldots, w_L\right] \\
&= \mathbb{E}_{h,T_L}\left[\mathbb{E}_{Z_L\sim D_L}[\ell(h, Z_L)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L_i})\right] .
\end{aligned}
\tag{22}
$$

We are now going to upper bound

$$\mathbb{E}_{h,T_L}\left[\mathbb{E}_{Z_L\sim D_L}[\ell(h, Z_L)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L_i})\right] . \tag{23}$$

Note that $T_L \sim D_L^n$ when give $w_1, \ldots, w_L$ because of the Markov property. We adopt the classical idea of ghost sample in statical learning theory. That is, we sample another n-tuple T_L':

$$T_L' = \{Z_1', \ldots, Z_L'\} \tag{24}$$

8

where each element Z'_i is drawn i.i.d from the distribution D_L. We now have,

$$\mathbb{E}_{h,T_L}\left[\mathbb{E}_{Z_L \sim D_L}[\ell(h, Z_L)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L_i})\right]$$

$$= \mathbb{E}_{h,T_L}\left[\mathbb{E}_{T'_L}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L'_i})\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L_i})\right]$$

$$= \mathbb{E}_{h,T_L,T'_L}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L'_i})\right] - \mathbb{E}_{h,T_L}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L_i})\right]. \tag{25}$$

We know that the output classifier in the output layer follows the distribution $P_{h|T_L}$. We denote the joint distribution of h and T_L by $P_{h,T_L} = P_{h|T_L} \times P_{T_L}$. Also, we denote the marginal distribution of h and T_L by P_h and P_{T_L} respectively. Therefore, we have,

$$\mathbb{E}_{h,T_L,T'_L}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L'_i})\right] - \mathbb{E}_{h,T_L}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L_i})\right]$$

$$= \mathbb{E}_{h' \sim P_h, T'_L \sim P_{T_L}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h', Z_{L'_i})\right] - \mathbb{E}_{(h,T_L) \sim P_{h,T_L}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_{L_i})\right]. \tag{26}$$

We now bound the above term by the mutual information $I(T_L, h)$ by employing the following lemma.

Lemma 3 (Donsker and Varadhan 1983). *Let P and Q be two probability distributions on the same measurable space $\{\Omega, \mathcal{F}\}$. Then the KL-divergence between P and Q can be represent as,*

$$D(P||Q) = \sup_{F}\left[\mathbb{E}_P[F] - \log\mathbb{E}_Q[e^F]\right] \tag{27}$$

where the supreme is taken over all measurable functions $F : \Omega \to \mathbb{R}^+$ such that $\mathbb{E}_Q[e^F] < \infty$.

Using lemma 3, we have,

$$I(T_L, h) = D(P_{h,T_L}||P_h \times P_{T_L})$$

$$= \sup_{F}\left[\mathbb{E}_{P_{h,T_L}}[F] - \log\mathbb{E}_{P_h \times P_{T_L}}\left[e^F\right]\right]$$

$$\geq \mathbb{E}_{(h,T_L) \sim P_{h,T_L}}\left[\frac{\lambda}{n}\sum_{i=1}^{n}\ell(h, Z_{L_i})\right] - \log\mathbb{E}_{h' \sim P_h, T'_L \sim P_{T_L}}\left[e^{\frac{\lambda}{n}\sum_{i=1}^{n}\ell(h', Z_{L'_i})}\right]. \tag{28}$$

As the loss function $\ell(h', Z_{L'_i})$ is σ-sub-Gaussian w.r.t $Z_{L'_i}$ for any $h' \sim P_h$ and $Z_{L'_i} \sim D_L$ is i.i.d for $i = 1, \ldots, n$, then $\frac{1}{n}\sum_{i=1}^{n}\ell(h', Z_{L'_i})$ is $\frac{\sigma}{\sqrt{n}}$-sub-Gaussian. By definition, we have,

$$\log\mathbb{E}_{h' \sim P_h, T'_L \sim P_{T_L}}\left[e^{\frac{\lambda}{n}\sum_{i=1}^{n}\ell(h', Z_{L'_i})}\right] \leq \frac{\sigma^2\lambda^2}{2n} + \mathbb{E}_{h' \sim P_h, T'_L \sim P_{T_L}}\left[\frac{\lambda}{n}\sum_{i=1}^{n}\ell(h', Z_{L'_i})\right]. \tag{29}$$

9

Substituting inequality (29) into inequality (28), we have,

$$
\mathbb{E}_{(h,T_L)\sim P_{h,T_L}} \left[\frac{\lambda}{n} \sum_{i=1}^{n} \ell(h, Z_{L_i}) \right] - \frac{\sigma^2 \lambda^2}{2n}
$$

$$
- \mathbb{E}_{h'\sim P_h, T'_L \sim P_{T_L}} \left[\frac{\lambda}{n} \sum_{i=1}^{n} \ell(h', Z_{L'_i}) \right] - I(T_L, h)
$$

$$
= -\frac{\sigma^2 \lambda^2}{2n} - I(T_L, h)
$$

$$
- \left[\mathbb{E}_{h'\sim P_h, T'_L \sim P_{T_L}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(h', Z_{L'_i}) \right] - \mathbb{E}_{(h,T_L)\sim P_{h,T_L}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_{L_i}) \right] \right] \lambda
$$

$$
\leq 0 . \tag{30}
$$

The above inequality is a quadratic curve about λ and always less than or equal to zero. Therefore we have,

$$
|\mathbb{E}[R(W) - R_S(W)|w_1, \ldots, w_L]|^2 \leq \frac{2\sigma^2}{n} I(T_L, h) \tag{31}
$$

which completes the proof.

$\square$

By Theorem 1, we can use the strong data processing inequality for a Markov chain in figure 2 recursively. Thus, we have,

$$
\sqrt{\frac{2\sigma^2}{n} I(T_L, h)} \leq \sqrt{\frac{2\sigma^2}{n} \eta_L I(T_{L-1}, h)}
$$

$$
\leq \sqrt{\frac{2\sigma^2}{n} \eta_L \eta_{L-1} I(T_{L-1}, h)}
$$

$$
\leq \ldots \leq \sqrt{\frac{2\sigma^2}{n} \left(\prod_{k=1}^{L} \eta_k \right) I(S, h)}
$$

$$
\leq \sqrt{\frac{2\sigma^2 \eta^L}{n} I(S, h)}
$$

$$
= \exp\left(-\frac{L}{2} \log \frac{1}{\eta} \right) \sqrt{\frac{2\sigma^2}{n} I(S, h)} \tag{32}
$$

where

$$
\eta = \max_{i \in \{1,\ldots,L\}} \eta_i < 1 . \tag{33}
$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}[R(W) - R_S(W)] \\
&= \mathbb{E}\left[\mathbb{E}[R(W) - R_S(W)|w_1,\ldots,w_L]\right] \\
&\leq \mathbb{E}\left(\exp\left(-\frac{L}{2}\log\frac{1}{\eta}\right)\sqrt{\frac{2\sigma^2}{n}I(S,h)}\right) \\
&= \exp\left(-\frac{L}{2}\log\frac{1}{\eta}\right)\sqrt{\frac{2\sigma^2}{n}I(S,h)} \\
&\leq \exp\left(-\frac{L}{2}\log\frac{1}{\eta}\right)\sqrt{\frac{2\sigma^2}{n}I(S,W)} \quad\quad (34)
\end{aligned}
$$

which completes the proof of Theorem 2.

6   Conclusions

In this paper, we obtain an exponential bound for the expected generalization error of deep learning algorithms. Our results have valuable implications to other critical problems in deep learning and need to be further investigated. (1) The traditional statistical learning theory can validate the success of deep neural networks because the mutual information $I(S,W)$ decreases with the increasing number of layers L and the fact that smaller mutual information implies higher algorithmic stability (Raginsky et al. 2016) and smaller complexity of the algorithmic hypothesis class (Liu et al. 2017). (2) The information loss factor $\eta < 1$ has the potential to study characteristics of various pooling and activation functions as well as deep learning tricks (how they contribute to the reduction of the expected generalization error).

References

Ahlswede, R. and Gács, P. (1976). Spreading of sets in product spaces and hypercontraction of the markov operator. *The annals of probability*, pages 925–939.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pages 231–238.

Donsker, M. D. and Varadhan, S. S. (1983). Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212.

Langford, J. and Shawe-Taylor, J. (2003). Pac-bayes & margins. In *Advances in neural information processing systems*, pages 439–446.

Liu, T., Lugosi, G., Neu, G., and Tao, D. (2017). Algorithmic stability and hypothesis complexity. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2159–2167. PMLR.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.

Polyanskiy, Y. and Wu, Y. (2015). Strong data-processing inequalities for channels and Bayesian networks. *ArXiv e-prints*.

Raginsky, M., Rakhlin, A., Tsao, M., Wu, Y., and Xu, A. (2016). Information-theoretic analysis of stability and bias of learning algorithms. In *Information Theory Workshop (ITW), 2016 IEEE*, pages 26–30. IEEE.

Russo, D. and Zou, J. (2015). How much does your data exploration overfit? Controlling bias via information usage. *ArXiv e-prints*.

Shwartz-Ziv, R. and Tishby, N. (2017). Opening the Black Box of Deep Neural Networks via Information. *ArXiv e-prints*.

Sonoda, S. and Murata, N. (2015). Neural network with unbounded activation functions is universal approximator. *arXiv preprint arXiv:1505.03654*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems 30*, pages 2524–2533. Curran Associates, Inc.

Xu, H. and Mannor, S. (2012). Robustness and generalization. *Machine learning*, 86(3):391–423.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550.