Hierarchical Generalization Bounds for Deep Neural Networks

Haiyun He
Center for Applied Mathematics
Cornell University
Ithaca, NY, USA
Email: hh743@cornell.edu

Christina Lee Yu
School of ORIE
Cornell University
Ithaca, NY, USA
Email: cleeyu@cornell.edu

Ziv Goldfeld
School of ECE
Cornell University
Ithaca, NY, USA
Email: goldfeld@cornell.edu

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619279

***Abstract*—Deep neural networks (DNNs) exhibit an exceptional generalization capability in practice. This work aims to capture the effect of depth and its potential benefit for learning within the paradigm of information-theoretic generalization bounds. We derive two novel hierarchical bounds on the generalization error that explicitly depend on the internal representations within each layer. The first result, is a layer-dependent generalization bound in terms of the Kullback-Leibler (KL) divergence, which shrinks as the layer index increases. The second bound, which is based on the Wasserstein distance, implies the existence of a layer that serves as a generalization funnel, which minimizes the generalization bound. We then specialize our bounds to the case of binary Gaussian classification, and present analytic expressions dependent on weight matrices rank or certain norms, for the KL divergence and the Wasserstein bounds, respectively. Our results may provide a new perspective for understanding generalization in deep models.**

I. Introduction

Overparameterized deep neural networks (DNNs) have surged in popularity as the preferred model for numerous high-dimensional and large-scale learning tasks, primarily due to their remarkable generalization performance. Substantial efforts have been devoted to theoretically explaining this phenomenon from various perspectives. This includes norm-based complexity measures [1]–[3], PAC-Bayes bounds [4]–[9], sharpness and flatness of the loss minima [10]–[12], loss landscape [13], implicit regularization induced by the gradient descent algorithms [14]–[16], etc. The reader is referred to the recent survey [17] for a comprehensive literature review. Despite this wealth of research, the precise factors contributing to the generalization capacity of DNNs remain elusive, as indicated in [18], [19]. The goal of this work is to shed new light on the advantages of deep models for learning under the framework of information-theoretic generalization bounds.

The generalization error is the difference between the population risk and the empirical risk on the training data. It measures the extent of overfitting of a trained neural network when the empirical risk is pushed to zero. Information-theoretic generalization bounds have been widely explored in recent years. This line of work was initiated by [20], where a generalization error bound in terms of the mutual information between the input and output of the learning algorithm was derived; see also [21], [22]. These inaugural results inspired various extensions and refinements based on chaining arguments [23], [24], conditioning and processing techniques [25]–[28], as well as other information-theoretic quantities [29]–[32]. However, the aforementioned results were not specialized to the DNN setting. Hence, they did not explicitly elucidate the impact of the structure of DNNs, including factors such as the number of layers, parameter size, and the used activation functions, on generalization. The paper [33] considered the multilayer structure of DNN, especially under the Gibbs algorithm, but did not address the dependence of generalization on the network architecture and parameters. Quantifying these effects within information-theoretic bounds is the main objective of this work.

Towards this goal, we present two new hierarchical generalization error bounds for DNNs. The first bound refines the results from [20]–[22], by bounding the generalization in terms of the Kullback-Leibler (KL) divergence and mutual information associated with the internal representations of each layer. This bound shrinks as the layer count increases, can adapt to layers of low complexity (e.g., low-dimensional or discrete), and overall highlights the benefits of depth for learning. Our second generalization bound explores an alternative approach by bounding the generalization in terms of the Wasserstein distance associated with the layer indices. This bound implies that there exists a layer that minimizes the generalization upper bound, which serves as a *generalization funnel layer*. To quantify these, we specialize our bounds to the case of binary Gaussian mixture classification problem. The derived analytic expressions show that as we delve deeper into the network, the KL divergence bounds shrink as a result of the shrinking ranks of the product of weight matrices; the generalization funnel layer induced by the Wasserstein bounds depends on the Frobenius norms of the weight matrices product. We compute the generalization funnel layer using a simple numerical example, which shows that the funnel layer depends on the model generating methods.

The rest of our paper is organized as follows. In Section II, we define the notations and formulate the supervised learning problem under a feedforward DNN model. In Section III present the hierarchical generalizaiton bounds based on the KL divergence and the Wasserstein distance, respectively. We then specialize our bounds to the case of binary Gaussian classification and derive the analytic expressions. We conclude

our discussion and present avenues for future work in Section IV. The proofs of our results are provided in [34, Appendix].

II. Preliminaries and Problem Formulation

A. Notation

The class of Borel probability measures on $\mathcal{X} \subseteq \mathbb{R}^d$ is denoted by $\mathcal{P}(\mathcal{X})$. A random variable $X \sim P_X \in \mathcal{P}(\mathcal{X})$ is called σ-sub-Gaussian, if $\mathbb{E}\big[\exp\big(\lambda(X-\mathbb{E}[X])\big)\big] \leq \lambda^2\sigma^2/2$ for any $\lambda \in \mathbb{R}$. The f-divergence between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ ($\mu \ll \nu$) is defined by $\mathsf{D}_f(\mu\|\nu) \coloneqq \int f(d\mu/d\nu)\, d\nu$, where $f : (0, +\infty) \to \mathbb{R}$ is convex and $f(1) = 0$. The Kullback-Leibler (KL) divergence is defined by taking $f(u) = u \log u$. The Hellinger (H^2) distance is defined by taking $f(u) = (1-\sqrt{u})^2$. The total variation (TV) distance is defined by taking $f(u) = \frac{1}{2}|u-1|$. The mutual information between $(X, Y) \sim P_{X,Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is defined as $\mathsf{I}(X;Y) \coloneqq \mathsf{D}_{\mathsf{KL}}(P_{X,Y}\|P_X \otimes P_Y)$. The Shannon entropy of a discrete random variable $X \sim P_X \in \mathcal{P}(\mathcal{X})$ is $\mathsf{H}(X) = \mathsf{H}(P_X) = \log(|\mathcal{X}|) - \mathsf{D}_{\mathsf{KL}}(P_X\|\mathrm{Unif}(\mathcal{X}))$. Suppose $\mathcal{X}$ is a complete separable metric space; for $p \in \mathbb{N}$ and $p \geq 1$, the p-Wasserstein distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is defined as $\mathsf{W}_p(\mu,\nu) \coloneqq (\inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(x,x')\sim\pi}[\|x-x'\|^p])^{1/p}$, where $\Pi(\mu,\nu)$ denotes the set of couplings on $\mathcal{X}^2$ with marginal distributions μ and ν. For a d-dimensional vector X and integers $1 \leq i < j \leq d$, we use the shorthands $X_i^j \coloneqq (X_i, \ldots, X_j)$ and $[j] \coloneqq \{1, 2, \ldots, j\}$. For a vector v, define $\|v\| \coloneqq \sqrt{v^\mathsf{T} v}$ as the Euclidean norm. For a matrix $\mathbf{A}$, define $\|\mathbf{A}\|_{\mathrm{op}} = \sup\{\|\mathbf{A}v\| \mid \|v\| = 1\}$ as the operator norm and $\|\mathbf{A}\|_{\mathrm{F}} \coloneqq \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^*)}$ as the Frobenius norm.

B. Supervised Learning Problem

Consider a data space $\mathcal{X} \subseteq \mathbb{R}^{d_0}$ and label set $\mathcal{Y} = [K] \subseteq \mathbb{N}$. Fix a data distribution $P_{X,Y} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ and let $(X, Y) \sim P_{X,Y}$ be a nominal data feature–label pair. The training dataset $D_n = \{(X_i, Y_i)\}_{i=1}^n$ comprises independently and identically distributed (i.i.d.) copies of (X, Y); note that $P_{D_n} = P_{X,Y}^{\otimes n}$. We consider a feedforward DNN model with L layers for predicting the label Y from the test sample X via $\hat{Y} \coloneqq g_{\mathbf{w}_L} \circ g_{\mathbf{w}_{L-1}} \circ \cdots \circ g_{\mathbf{w}_1}(X)$, where $g_{\mathbf{w}_l}(t) = \phi_l(\mathbf{w}_l t)$, $l \in [L]$, for a weight matrix $\mathbf{w}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ and an activation function $\phi_l : \mathbb{R} \to \mathbb{R}$ (acting on vectors element-wise). Denote all the network parameters by $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_L)$ and the parameter space by $\mathcal{W} \subseteq \mathbb{R}^{d_1\times d_0} \times \cdots \times \mathbb{R}^{d_L \times d_{L-1}}$. We denote the internal representation of the l^{th} layer by $T_l \coloneqq g_{\mathbf{w}_l} \circ \cdots \circ g_{\mathbf{w}_1}(X)$, $l \in [L]$, noting that $T_0 = X$. When the input to the network is X_i (rather than X), we add a subscript i to the internal representation notation, writing $T_{l,i}$ instead of T_l. See Figure 1 for an illustration. We know that the setup can be generalized to regression problems by setting $\mathcal{Y} \subseteq \mathbb{R}$. Furthermore, our arguments extend to the case when the training dataset D_n comprises dependent but identically distributed data samples, e.g., ones generated from a Markov chain Monte Carlo method.

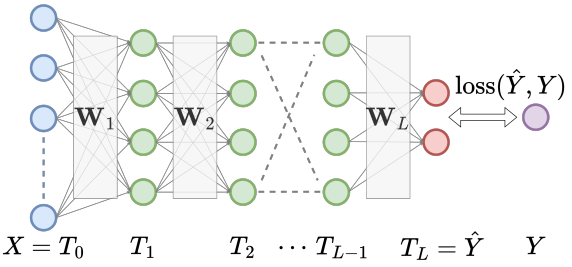

Fig. 1. L-layer feedforward network.

Let $\ell : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be the loss function. Given any $\mathbf{w} \in \mathcal{W}$, the *population risk* and the *empirical risk* are respectively defined as

$$\mathcal{L}_{\mathsf{P}}(\mathbf{w}, P_{X,Y}) \coloneqq \mathbb{E}[\ell(\mathbf{w}, X, Y)];$$
$$\mathcal{L}_{\mathsf{E}}(\mathbf{w}, D_n) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}, X_i, Y_i),$$

where the loss function ℓ penalizes the discrepancy between the true label Y and the DNN prediction $\hat{Y} = g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(X)$, i.e., $\ell(\mathbf{w}, x, y) = \tilde{\ell}(g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(x), y)$. A learning algorithm trained with D_n can be characterized by a stochastic mapping $P_{\mathbf{W}|D_n}$. Given any $(P_{\mathbf{W}|D_n}, P_{X,Y})$, the *expected generalization error* is defined as the expected gap between the population empirical risks:

$$\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) \coloneqq \mathbb{E}[\mathcal{L}_{\mathsf{P}}(\mathbf{W}, P_{X,Y}) - \mathcal{L}_{\mathsf{E}}(\mathbf{W}, D_n)], \quad (1)$$

where the expectation is w.r.t. $P_{(X,Y),D_n,\mathbf{W}} = P_{X,Y}^{\otimes(n+1)} \otimes P_{\mathbf{W}|D_n}$.

III. Hierarchical Generalization Bound

Existing results such as [20], [22] bound the generalization error from (1) in terms of the mutual information terms $\mathsf{I}(D_n;\mathbf{W})$ or $\sum_{i=1}^n \mathsf{I}(X_i, Y_i;\mathbf{W})$, which only depend on the raw input dataset and the algorithm. We next establish two new improved generalization bounds, whose hierarchical structure captures the effect of the internal representations T_l. Notably, the first bound shrinks as one moves deeper into the network, providing new evidence for the benefits of deep models for learning. The second bound is minimized by one of the network layers, shedding light on understanding the different effects of internal representations.

A. KL Divergence Bound

We present the following generalization bound for the above described setting.

Theorem 1 (Hierarchical generalization bound)**.** *Suppose that the loss function $\ell(\mathbf{w}, X, Y)$ is σ-sub-Gaussian under $P_{X,Y}$, for all $\mathbf{w} \in \mathcal{W}$. We have*

$$\big|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})\big| \leq \mathsf{UB}(L) \leq \mathsf{UB}(L-1) \leq \ldots \leq \mathsf{UB}(0),$$

where $\mathsf{UB}(l) \coloneqq \frac{\sigma\sqrt{2}}{n}\sum_{i=1}^n \big(\mathsf{I}(T_{l,i}, Y_i; \mathbf{W}_{l+1}^L|\mathbf{W}_1^l) + \mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\big\|P_{T_l,Y|\mathbf{W}_1^l}\big|P_{\mathbf{W}_1^l})\big)^{1/2}$, $l = 0, \ldots, L$.

Theorem 1 is derived by first establishing the $\mathsf{UB}(L)$ upper bound via the Donsker-Varadhan variational representation of the KL divergence and the sub-Gaussianity of the loss function. We then invoke the data processing inequality (DPI) to successively peel off the layers to arrive at the remaining bounds. See [34, Appendix A] for a detailed proof. While the $\mathsf{UB}(L)$ forms the tightest bound, the state hierarchy highlights the benefit of depth for learning and lend well for comparison to existing results. Indeed, observing that $\mathsf{UB}(0) = \sqrt{2\sigma^2}\, n^{-1} \sum_{i=1}^{n} \sqrt{\mathsf{I}(X_i, Y_i; \overline{\mathbf{W}})}$, we see that our bound is indeed tighter than the one from [22].

Theorem 1 shows that the model generalizes when both $\mathsf{I}(T_{l,i}, Y_i; \mathbf{W}_{l+1}^{L} | \mathbf{W}_1^l)$ and $\mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l} \| P_{T_l,Y|\mathbf{W}_1^l} | P_{\mathbf{W}_1^l})$ are small, for some layer $l = 0, \ldots, L$. This happens when the weights of subsequent layers are not overly dependent on the l^{th} input internal representation, and when the learned posterior of this internal representation highly matches the prior.

Special case (discrete latent space). When T_l only takes a finite number of values, i.e., its support satisfies $|\mathcal{T}_l| < \infty$ (e.g., the discrete latent layer in the VQ-VAE [35]). Assuming that $t_l(\mathbf{w}_1^l) \coloneqq \min_{t \in \mathcal{T}_l, y \in \mathcal{Y}} P_{T_l,Y|\mathbf{W}_1^l}(t, y | \mathbf{w}_1^l) \in \left(0, |\mathcal{T}_l \times \mathcal{Y}|^{-1}\right)$ and $\overline{t_l} \coloneqq \sup_{\mathbf{w}_1^l} t_l(\mathbf{w}_1^l)$, we have

$$\mathsf{UB}(l) \leq \sqrt{2\sigma^2 \log\left(K^2/\overline{t_l}\right)}.$$

As $\overline{t_l}$ grows, we see that $P_{T_l,Y|\mathbf{W}_1^l}$ tends to the uniform distribution on $T_l \times \mathcal{Y}$ and its entropy/variance increases. This, in turn, shrinks the generalization error, which is consistent with the intuition that stochasticity leads to better generalization.

Proof Sketch: The information measures in $\mathsf{UB}(l)$ can be upper bounded as follows: $\mathsf{I}(T_{l,i}, Y_i; \mathbf{W}_{l+1}^{L} | \mathbf{W}_1^l) + \mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l} \| P_{T_l,Y|\mathbf{W}_1^l} | P_{\mathbf{W}_1^l}) \leq 2\mathsf{H}(Y_i | T_{l,i}, \mathbf{W}_1^l) - \mathbb{E}_{P_{T_{l,i},Y_i,\mathbf{W}_1^l}}[\log P_{T_l,Y|\mathbf{W}_1^l}] \leq 2\log K - \mathbb{E}_{P_{T_{l,i},Y_i,\mathbf{W}_1^l}}[\log P_{T_l,Y|\mathbf{W}_1^l}] \leq \log(K^2/\overline{t_l})$.

B. Wasserstein Distance Bound

Akin to Theorem 1, we present a generalization error bound based on the Wasserstein distance. Unlike the KL divergence, Wasserstein distances do not generally follow the DPI, and hence the presented bound does not adhere to a descending hierarchical structure. Instead, it shows that there exists a layer that minimizes the Wasserstein generalization bound.

Theorem 2 (Min Wasserstein generalization bound)**.** *Suppose that the loss function $\tilde{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is ρ_0-Lipschitz and the activation function $\phi_l : \mathbb{R} \to \mathbb{R}$ is ρ_l-Lipschitz, for each $l = 1, \ldots, L$. We have*

$$\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) \leq \min_{l=0,\ldots,L} \frac{\rho_0}{n} \sum_{i=1}^{n} \mathbb{E}\Bigg[\left(1 \vee \prod_{j=l+1}^{L} \rho_j \|\mathbf{W}_j\|_{\text{op}}\right) \mathsf{W}_1\big(P_{T_{l,i},Y_i|\mathbf{W}}(\cdot|\mathbf{W}), P_{T_l,Y|\mathbf{W}}(\cdot|\mathbf{W})\big)\Bigg].$$

The derivation of the bound relies on Kantorovich–Rubinstein duality, which ties W_1 to the difference of expectations defining the generalization error. See [34, Appendix B] for the proof details. As the Wasserstein distance is monotonically increasing in the order (i.e., $\mathsf{W}_p \leq \mathsf{W}_q$ whenever $p \leq q$), the 1-Wasserstein distance provides the sharpest bound. Compared to the KL divergence bound from Theorem 1, which degenerates when the considered distributions are supported on different domains, the Wasserstein distance is robust to mismatched supports and the corresponding bound is meaningful even in that setting.

Theorem 2 suggests that the generalization bound is controlled by a certain layer that achieves the smallest weighted 1-Wasserstein distance between the distributions of the training and test internal representations. This layer serves as a funnel that determines the overall generalization performance; thus, we call it *generalization funnel layer*. It suggests that within a DNN, there exists a specific layer that exerts a stronger impact on generalization compared to others.

Remark 1 (Comparison with KL-divergence based bound)**.** *Assume that the loss function (bounded within $[0, A] \subset \mathbb{R}_{\geq 0}$) and the activation functions in the DNN model satisfy the Lipschitz continuity conditions in Theorem 2. Under this assumption, the loss function $\ell(\mathbf{w}, X, Y)$, where $(X, Y) \sim P_{X,Y}$, is $\frac{A}{2}$-subGaussian for all $\mathbf{w}$. When $\rho_0 K^2 \leq A$, the generalization bound given in Theorem 2 is tighter than $\mathsf{UB}(L)$ in Theorem 1. A proof of this claim is provided in [34, Appendix D] , and utilizes [36, Theorem 4], Pinsker's and Bretagnolle-Huber inequalities.*

C. Case Study: Binary Gaussian Mixture Classification

To better understand the generalization bounds from Theorems 1 and 2 and assess their dependence on depth, we consider the following binary Gaussian mixture example and evaluate the bounds analytically.

Classification problem setting. Consider the binary classification problem illustrated in Fig. 2, where the input data distribution is a binary Gaussian mixture: $P_Y = \text{Unif}\{-1, +1\}$ and $P_{X|Y=y} = \mathcal{N}(y\mu_0, \sigma_0^2 \mathbf{I}_{d_0})$, where $\mu_0 \in \mathbb{R}^d$ and $\sigma_0 > 0$. The goal is to classify the binary label Y given the feature X. Notice that under this setting, the Bayes optimal linear classifier is $Y^\star = \tanh(\mu_0^\intercal X)$.

Model and algorithm. Consider a classifier that is realized by a linear L-layer neural network composed with a hyperbolic tangent nonlinearity at the output, i.e., $\hat{Y}(\mathbf{w}) = \tanh(\mathbf{w}_{\otimes L} X)$, where $\mathbf{w}_{\otimes l} \coloneqq \mathbf{w}_l \mathbf{w}_{l-1} \cdots \mathbf{w}_1$. To train the model to approaches the Bayes optimal classifier $\tanh(\mu_0^\intercal X)$, we consider an algorithm $P_{\mathbf{W}|D_n}$ defined by $\mathbf{W}_{\otimes L}^\intercal = \frac{1}{n} \sum_{i=1}^{n} Y_i X_i$, and set the prediction to $\hat{Y} = \hat{Y}(\mathbf{W})$. The rationale behind this choice of algorithm comes from observing that there are i.i.d. $Y_i X_i \sim \mathcal{N}(\mu_0, \sigma_0^2 \mathbf{I}_{d_0})$, for $i = 1, \ldots, n$. Consequently, $\mathbf{W}_{\otimes L}^\intercal$ can be viewed as the sample mean of the dataset $\{X_1 Y_1, \ldots, X_n Y_n\}$, and by the strong law of large number we have $\mathbf{W}_{\otimes L}^\intercal \to \mu_0$ almost surely, as $n \to \infty$. Performance is measured using the quadratic loss function $\ell(\mathbf{w}, X, Y) = \big(Y - \tanh(\mathbf{w}_{\otimes L} X)\big)^2$, which is bounded inside $[0, 4]$ and is thus 2-sub-Gaussian under $P_{X,Y}$, for all $\mathbf{w}$.

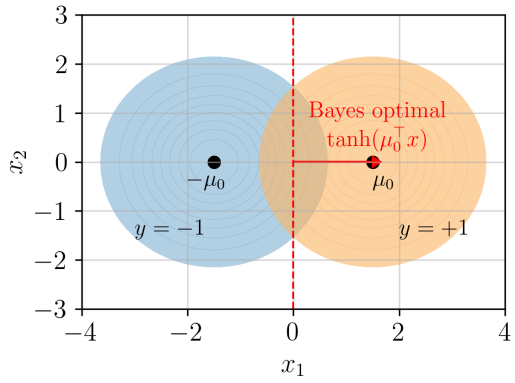

Fig. 2. Binary Gaussian mixture data when $d_0 = 2$.

Analysis. We move to evaluate the generalization bounds in Theorems 1 and 2 by computing the prior and posterior distributions and the divergences between them. Proofs of subsequent claims are all deferred to [34, Appendix E].

Lemma 3 (Prior and posterior of (X_i, Y_i))**.** *For any $i \in [n]$ and $y \in \{\pm 1\}$, the prior distribution of $X_i|Y_i = y$ is given by $P_{X_i|Y_i=y} = \frac{1}{2}\mathcal{N}(y\mu_0, \sigma_0^2 \mathbf{I}_{d_0})$, while its posterior distribution given the model $\mathbf{W}_{\otimes L}$ is $P_{X_i|Y_i=y,\mathbf{W}_{\otimes L}} = \frac{1}{2}\mathcal{N}(y\mathbf{W}_{\otimes L}^\mathsf{T}, \frac{(n-1)\sigma_0^2}{n}\mathbf{I}_{d_0})$. Furthermore, we have $P_{Y_i|\mathbf{W},\mathbf{W}_{\otimes L}} = P_{Y_i|\mathbf{W}_{\otimes L}} = \mathrm{Unif}\{-1,+1\}$ and $P_{X_i,Y_i|\mathbf{W},\mathbf{W}_{\otimes L}} = P_{X_i,Y_i|\mathbf{W}_{\otimes L}}$.*

Given the above expressions for the involved distributions, we evaluate the KL divergence generalization bound from Theorem 1 as follows.

Proposition 4 (KL divergence bound evaluation)**.** *Under the binary Gaussian classification setting, we have*

$$\left|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})\right| \le \widetilde{\mathsf{UB}}(L) \le \widetilde{\mathsf{UB}}(L-1) \le \cdots \le \widetilde{\mathsf{UB}}(0),$$

where $\widetilde{\mathsf{UB}}(l) \coloneqq 2\sqrt{r_l(\log\frac{n}{n-1} - \frac{1}{n}) + \frac{d_0}{n}}$, $r_0 = d_0$, and $r_l = \mathrm{rank}(\mathbf{W}_{\otimes l})$, for $l \in [L]$.

As a sanity check, observe that $\widetilde{\mathsf{UB}}_n(l)$ converges to 0 as $n \to \infty$, for all $l = 0, 1, \ldots, L$, as expected. Recalling that $\mathrm{rank}(\mathbf{AB}) \le \mathrm{rank}(\mathbf{A}) \wedge \mathrm{rank}(\mathbf{B})$, we see that $r_L \le r_{L-1} \le \cdots \le r_1 \le r_0 = d_0$. Consequently, the contraction from $\widetilde{\mathsf{UB}}_n(l-1)$ to $\widetilde{\mathsf{UB}}_n(l)$ is evident and quantified by the gap between the ranks of $\mathbf{W}_{\otimes(l-1)}$ and $\mathbf{W}_{\otimes l}$, namely, $r_{l-1} - r_l$. Note that in our example, $\mathrm{rank}(\mathbf{W}_L) = 1$ and thus $\mathrm{rank}(\mathbf{W}_{\otimes L}) = 1$, independent of the depth L, which means that the tightest bound, $\widetilde{\mathsf{UB}}(L)$, does not change with L. Nevertheless, the intermediate bounds $\widetilde{\mathsf{UB}}(l)$, for $l \in [L-1]$, generally shrink as L grows, depicting the trajectory of generalization performance of the internal layers. Extending the above example beyond the classification setting to representation learning, where the output representation dimension d_L varies according to the network structure, would enable observing a similar effect for $\widetilde{\mathsf{UB}}(L)$ as well. Our focus on the binary classification case is motivated by its analytic tractability, and we leave further extensions for future work.

TABLE I

THE GENERALIZATION FUNNEL LAYER INDEX l^* FOR DIFFERENTLY GENERATED MODEL $\mathbf{W}$ WHEN $L = 10$ IN EXAMPLE 1. THE GENERATING METHOD IS DETERMINED BY $\prod_{j=1}^{l} C_j = 0.2\|\frac{1}{n}\sum_{i=1} Y_i X_i\|$.

Generating method	$l' = 3$	$l' = 5$	$l' = 7$
Generalization funnel layer l^*	3	5	7

We proceed to evaluate the Wasserstein generalization bound under the considered setting.

Proposition 5 (Min Wasserstein distance bound evaluation)**.** *Under the binary Gaussian classification setting and from Theorem 2, we have*

$$\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) \le \min_{l=0,\ldots,L} \mathsf{WUB}(l),$$

where $\mathsf{WUB}(l) \coloneqq (4\sqrt{2}\sigma_0(\sqrt{d_0} + (\sqrt{n} - \sqrt{n-1}))/\sqrt{n})\mathbb{E}[(1 \vee \prod_{j=l+1}^{L}\|\mathbf{W}_j\|_{\mathrm{op}}^2)\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}^2]^{\frac{1}{2}}$, $\mathbf{W}_{\otimes l} = \mathbf{W}_l\mathbf{W}_{l-1}\cdots\mathbf{W}_1$ for $l \in [L]$, and $\mathbf{W}_{\otimes 0} = \mathbf{I}_{d_0}$.

Note that this upper bound also vanishes as $n \to \infty$. In this case, the *generalization funnel* layer that yields the tightest upper bound depends on the Frobenius norm of the product of network weight matrices up to the current layer $\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}}$ and the product of subsequent layers' operator norms. We notice that $\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\mathbf{W}_{\otimes l}\mathbf{W}_{\otimes l}^\mathsf{T})}$ not only depends on $\mathrm{rank}(\mathbf{W}_{\otimes l})$ but also on the singular values of $\mathbf{W}_{\otimes l}$. Thus, the generalization funnel layer is not necessarily the last one. In the following example, by considering a simple neural network model with different training methods, we empirically show that the generalization funnel layer depends on the training method.

Example 1 (Numerical evaluation of Proposition 5)**.** *Let $L = 10$, $d_0 = d_1 = \cdots = d_{L-1} = 2$, $n = 100$, $\mu_0 = (0.5, 0)$, $\sigma_0 = 1$. We generate the network model parameters as follows: $\mathbf{W}_L = (0, C_L)$, $\mathbf{W}_l = C_l\begin{pmatrix}\cos\theta_l & \sin\theta_l \\ -\sin\theta_l & \cos\theta_l\end{pmatrix}$ for $l = 1, \ldots, L-1$, where $\sum_{l=1}^{L-1}\theta_l = \arccos\langle \mathbf{W}_L, \frac{1}{n}\sum_{i=1} Y_i X_i\rangle$ and $\prod_{l=1}^{L} C_l = \|\frac{1}{n}\sum_{i=1} Y_i X_i\|$. $(\mathbf{W}_1, \ldots, \mathbf{W}_{L-1})$ are scaled rotation matrices that rotate $\mathbf{W}_L$ to $\frac{1}{n}\sum_{i=1} Y_i X_i$. Under this model, we have that $\mathbf{W}_{\otimes l}$ is full-rank, $\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}} = \sqrt{2}\prod_{j=1}^{l} C_j$ for $l = 1, \ldots, L-1$, and $\|(\mathbf{W}_{\otimes L})_1^{r_l}\|_{\mathrm{F}} = \|\frac{1}{n}\sum_{i=1} Y_i X_i\|$. Given the training dataset D_n, we generate $\{C_l\}_{l=1}^{L}$ such that $\prod_{j=l+1}^{L} C_j \le 1$ for all $l = 0, 1, \ldots, L-1$ and $\prod_{j=1}^{l'} C_j$ be sufficiently small for $l' \in \{3, 5, 7\}$. We compute the generalization funnel layer index as the minimizer of the sample mean from 10^4 output network parameters $\mathbf{W}$ (trained on 100 datasets D_n): $l^* = \arg\min_{l=0,1,\ldots,L}$ SampleMean$(\|\mathbf{W}_{\otimes l}\|_{\mathrm{F}})$. As shown in Table I, the generalization funnel layer varies according to the parameter generating methods.*

IV. Conclusion and Future Work

This work has taken a novel step at understanding the generalization performance of DNNs from the perspective of information-theoretic generalization bounds. We built upon the existing information-theoretic results by specializing them to the DNN setting. We derived two hierarchical generalization bounds that capture the effect of depth through the internal representations of the corresponding layers. The two bounds compare the distributions of internal representations of the training and test data under (i) the KL divergence, and (ii) the 1-Wasserstein distance. The KL divergence bound diminishes as the layer index increases, indicating the advantage of deep network architectures. The Wasserstein bound is minimized by the so-called *generalization funnel layer*, providing new a insight that certain layers play a more prominent role than others in governing generalization performance. We instantiated these results to a binary Gaussian mixture classification task with linear DNNs. Simple analytic expressions for the two generalization bounds we obtained, with the KL divergence reducing to depend on (and shrinks with) the rank of the product of weight matrices, while the Wasserstein bound simplified to depend on the operation and Frobenius norms of the weight matrix product. The latter further implied that the generalization funnel layer of a given model varies with different training methods.

In the future, to draw more compelling conclusions, it is necessary to conduct thorough analyses and experiments for general algorithms/architectures beyond the binary Gaussian classification. It would also be interesting and insightful to quantify the contraction in the hierarchical bounds in terms of the DNN architecture parameters.

Acknowledgment

H. He is supported by Cornell CAM postdoctoral fellowship. C. Yu is partially supported by NSF grants CNS-1955997, AFOSR grant FA9550-23-1-0301, and by an Intel Rising Stars award. Z. Goldfeld is partially supported by NSF grants CCF-2046018, DMS-2210368, and CCF-2308446, and the IBM Academic Award.

References

[1] N. Golowich, A. Rakhlin, and O. Shamir, "Size-independent sample complexity of neural networks," in *Conference On Learning Theory*. PMLR, 2018, pp. 297–299.

[2] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, "Exploring generalization in deep learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5949–5958.

[3] T. Liang, T. Poggio, A. Rakhlin, and J. Stokes, "Fisher-Rao metric, geometry, and complexity of neural networks," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 888–896.

[4] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach," in *International Conference on Machine Learning*. PMLR, 2018, pp. 254–263.

[5] G. K. Dziugaite and D. M. Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," in *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017*. AUAI Press, 2017.

[6] D. A. McAllester, "Some Pac-Bayesian theorems," in *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 1998, pp. 230–234.

[7] ——, "PAC-Bayesian model averaging," in *Proceedings of the 12th Annual Conference on Computational Learning Theory*, 1999, pp. 164–170.

[8] B. Neyshabur, S. Bhojanapalli, and N. Srebro, "A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks," in *International Conference on Learning Representations*, 2018.

[9] W. Zhou, V. Veitch, M. Austern, R. P. Adams, and P. Orbanz, "Nonvacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach," in *International Conference on Learning Representations*, 2018.

[10] S. Hochreiter and J. Schmidhuber, "Flat minima," *Neural Computation*, vol. 9, no. 1, pp. 1–42, 1997.

[11] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, "Sharp minima can generalize for deep nets," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1019–1028.

[12] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *International Conference on Learning Representations*, 2017.

[13] L. Wu, Z. Zhu, and W. E, "Towards understanding generalization of deep learning: Perspective of loss landscapes," *ICML 2017 Workshop on Principled Approaches to Deep Learning, Sydney, Australia*, 2017.

[14] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[15] S. L. Smith and Q. V. Le, "A Bayesian perspective on generalization and stochastic gradient descent," in *International Conference on Learning Representations*, 2018.

[16] S. Chatterjee and P. Zielinski, "On the generalization mystery in deep learning," *arXiv preprint arXiv:2203.10036*, 2022.

[17] D. Jakubovitz, R. Giryes, and M. R. Rodrigues, "Generalization error in deep learning," in *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*. Springer, 2019, pp. 153–193.

[18] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[19] K. Kawaguchi, L. P. Kaelbling, and Y. Bengio, "Generalization in deep learning," in *Mathematical Aspects of Deep Learning*. Cambridge University Press, 2022.

[20] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[21] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, 2019.

[22] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, 2020.

[23] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," in *Advances in Neural Information Processing Systems*, 2018, pp. 7234–7243.

[24] E. Clerico, A. Shidani, G. Deligiannidis, and A. Doucet, "Chained generalisation bounds," in *Conference on Learning Theory*. PMLR, 2022, pp. 4212–4257.

[25] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, "Conditioning and processing: Techniques to improve information-theoretic generalization bounds," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[26] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite, "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms." *Advances in Neural Information Processing Systems*, 2020.

[27] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Conference on Learning Theory*. PMLR, 2020, pp. 3437–3452.

[28] H. Harutyunyan, M. Raginsky, G. Ver Steeg, and A. Galstyan, "Information-theoretic generalization bounds for black-box learning algorithms," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 670–24 682, 2021.

[29] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-divergences and maximal leakage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, 2021.

[30] G. Aminian, S. Masiha, L. Toni, and M. R. Rodrigues, "Learning algorithm generalization error bounds via auxiliary distributions," *arXiv preprint arXiv:2210.00483*, 2022.
[31] G. Aminian, L. Toni, and M. R. Rodrigues, "Information-theoretic bounds on the moments of the generalization error of learning algorithms," in *IEEE International Symposium on Information Theory (ISIT)*, 2021.
[32] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via Wasserstein distance," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 577–581.
[33] A. R. Asadi and E. Abbe, "Chaining meets chain rule: Multilevel entropic regularization and training of neural networks," *Journal of Machine Learning Research*, vol. 21, no. 139, pp. 1–32, 2020.
[34] H. He, C. L. Yu, and Z. Goldfeld, "Information-theoretic generalization bounds for deep neural networks," *arXiv preprint arXiv:2404.03176*, 2024.
[35] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6306–6315, 2017.
[36] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.
[37] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2021, vol. 58.
[38] H. Wang, R. Gao, and F. P. Calmon, "Generalization bounds for noisy iterative algorithms using properties of additive noise channels," *Journal of machine learning research*, vol. 24, pp. 26:1–26:43, 2023.