Information-Theoretic Generalization Bounds for Deep Neural Networks

Haiyun He
Center for Applied Mathematics
Cornell University
hh743@cornell.edu

Christina Lee Yu
Department of ORIE
Cornell University
cleeyu@cornell.edu

Ziv Goldfeld
Department of ECE
Cornell University
goldfeld@cornell.edu

Abstract

Deep neural networks (DNNs) exhibit an exceptional capacity for generalization in practical applications. This work aims to capture the effect and benefits of depth for learning within the paradigm of information-theoretic generalization bounds. We derive two novel hierarchical bounds on the generalization error that capture the effect of the internal representations within each layer. The first bound demonstrates that the generalization bound shrinks as the layer index of the internal representation increases. The second bound aims to quantify the contraction of the relevant information measures when moving deeper into the network. To achieve this, we leverage the strong data processing inequality (SDPI) and employ a stochastic approximation of the DNN model we can explicitly control the SDPI coefficient. These results provide a new perspective for understanding generalization in deep models.

1 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 [17, 28, 25], PAC-Bayes bounds [3, 13, 26, 27, 29, 41], sharpness and flatness of the loss minima [21, 11, 24], loss landscape [38], implicit regularization induced by the gradient descent algorithms [34, 33, 7], etc. The reader is referred to the recent survey [22] 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 [40, 23]. 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 [39], where a generalization error bound in terms of the mutual information between the input and output of the learning algorithm was derived; see also [32, 6]. These inaugural results inspired various extensions and refinements based on chaining arguments [4, 8], conditioning and processing techniques [18, 19, 35, 20], as well as other information-theoretic quantities [14, 1, 2, 37]. However, the aforementioned results were not specialized to the DNN setting and hence did not capture the effect of depth on the generalization bound. Quantifying this effect within such 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 [6, 32, 39], by bounding the generalization in terms of information

Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

measures 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 is tailored to capture the contraction of the relevant information measures as we delve deeper into the network. To quantify this, we adopt a noisy DNN model (cf., e.g., [16, 15]) as a proxy for the deterministic network and employ the strong data processing inequality (SDPI) for the analysis. The SDPI coefficient associated with the stochastic channel induced by each layer is then controlled in terms of the layer dimension and the activation functions. The desired generalization bound is obtained by peeling off the DNN layers and aggregating the corresponding contraction coefficients. We visualize our bounds using a simple numerical example and discuss future research avenues stemming from our results.

2 Preliminaries and Problem Formulation

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 $\mu, \nu \in \mathcal{P}(\mathcal{X})$ ($\mu \ll \nu$[1]) 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 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 $\mathsf{H}(X) = \log(|\mathcal{X}|) - \mathsf{D}_{\mathsf{KL}}(P_X\|\mathsf{Unif}(\mathcal{X}))$. 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 function $f : \mathbb{R}^d \to \mathbb{R}^{d'}$, where $f = (f_1, \ldots, f_{d'})$, we define $\|f\|_\infty \coloneqq \sup_{x\in\mathbb{R}^d} \max_{i=1,\ldots,d'} |f_i(x)|$.

Supervised learning problem. Consider a data space $\mathcal{X} \subseteq \mathbb{R}^{d_0}$ and label set $\mathcal{Y} = [K] \subseteq \mathbb{Z}_+$. 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.

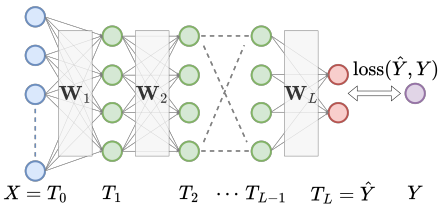

Figure 1: L-layer feedforward network.

Let $\ell : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_0^+$ be the loss function. Given any $\mathbf{w} \in \mathcal{W}$, the *population risk* and the *empirical risk* are respectively defined as

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

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., $\ell(\mathbf{w}, x, y) = \ell(g_{\mathbf{w}_L} \circ \cdots \circ g_{\mathbf{w}_1}(x), y)$. 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:

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

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

3 Generalization Error Bound via DPI

Existing results such as [39, 6] 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

[1] $\mu \ll \nu$ means μ is absolutely continuous w.r.t. ν.

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 bounds shrinks as one moves deeper into the network, providing new evidence for the benefits of deep models for learning.

3.1 Hierarchical Generalization Bound

Consider the setting described above, for which we present the following generalization error bound.

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*

$$|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \leq \mathsf{UB}(L) \leq \mathsf{UB}(L-1) \leq \ldots \leq \mathsf{UB}(0), \tag{2}$$

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

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, as it acts on the last layer T_L. We then invoke the data processing inequality (DPI) to successively peel off the layers to arrive at the remaining bounds. See 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 $\mathsf{UB}(0) = \sqrt{2\sigma^2}\, n^{-1} \sum_{i=1}^{n} \sqrt{\mathsf{I}(X_i, Y_i; \mathbf{W})}$, we see that our bound is indeed tighter than the one from [6].

Remark 1 (Interpretation and special cases)**.** *Theorem 1 shows that the model generalizes when, for some layer $l = 0, \ldots, L$, 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. This happens when the parameters of subsequent layers are not overly dependent on the l^{th} internal representation, and when the learned posterior of this internal representation highly matches the prior. To gain further intuition, we present two special cases under which the bounds simplify:*

1. *<u>One-to-one mapping:</u> If $g_{\mathbf{W}_l}$ is one-to-one for all $l \in [L]$, the DPI holds with equality:*

$$\mathsf{D}_{\mathsf{KL}}\big(P_{T_{l,i},Y_i|\mathbf{W}_1^l} \big\| P_{T_l,Y|\mathbf{W}_1^l} | P_{\mathbf{W}_1^l}\big) = \mathsf{D}_{\mathsf{KL}}\big(P_{X_i,Y_i|\mathbf{W}_1^l} \big\| P_{X,Y} | P_{\mathbf{W}_1^l}\big) = \mathsf{I}\big(X_i, Y_i; \mathbf{W}_1^l\big),$$
$$\mathsf{I}\big(T_{l,i}, Y_i; \mathbf{W}_{l+1}^{L} \big| \mathbf{W}_1^l\big) = \mathsf{I}\big(X_i, Y_i; \mathbf{W}_{l+1}^{L} \big| \mathbf{W}_1^l\big).$$

   Thus, the upper bounds are equal: $\mathsf{UB}(L) = \mathsf{UB}(L-1) = \cdots = \mathsf{UB}(0)$, which implies that the representation at each layer has the same impact on the generalization error.
2. *<u>Discrete latent space:</u> 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 [36]). Assuming that $t_l(\mathbf{w}_1^l) := \min_{t \in \mathcal{T}_l, y \in \mathcal{Y}} P_{T_l,Y|\mathbf{W}_1^l}(t, y | \mathbf{w}_1^l) \in \big(0, |\mathcal{T}_l \times \mathcal{Y}|^{-1}\big)$ and $\underline{t_l} := \sup_{\mathbf{w}_1^l} t_l(\mathbf{w}_1^l)$, we have*

$$\mathsf{UB}(l) \leq \sqrt{2\sigma^2 \log\big(K^2/\underline{t_l}\big)}.$$

   The proof is provided in Appendix B. As $\underline{t_l}$ 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.

3.2 Tighter Bound via Contraction for Noisy Networks

Inspired by Theorem 1, we next aim to capture the contraction of the information measures in our bound as the layer count grows. To that end, we use the *strong* DPI (SDPI) [10, 12, 31], for which some preliminaries are needed.

Strong data processing inequality. Given $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ and a transition kernel (channel) $P_{Y|X}$, write $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = P_{Y|X} \circ Q_X$ for the marginal distributions at the output of the channel when we feed it with P_X or Q_X, respectively. Assuming $P_X \ll Q_X$ and that Q_X is not a point mass, the SDPI coefficient for $P_{Y|X}$ under the f-divergence D_f is

$$\eta_f(P_{Y|X}) := \sup_{P_X, Q_X} \frac{\mathsf{D}_f(P_{Y|X} \circ P_X \| P_{Y|X} \circ Q_X)}{\mathsf{D}_f(P_X \| Q_X)} \in [0, 1].$$

We write $\eta_{\mathsf{KL}}(P_{Y|X})$ and $\eta_{\mathsf{TV}}(P_{Y|X})$ for the coefficients under the KL divergence and the total variation distance, respectively. Proposition II.4.10 in [9] shows that for any f-divergence, we have

$$\eta_f(P_{Y|X}) \leq \eta_{\mathsf{TV}}(P_{Y|X}) = \sup_{x, x' \in \mathcal{X}} \|P_{Y|X=x} - P_{Y|X=x'}\|_{\mathsf{TV}},$$

where $\eta_{\mathsf{TV}}(P_{Y|X})$ is also know as *Dobrushin's coefficient* [12]. It can be shown that if $Y = g(X)$ for some deterministic function $g : \mathcal{X} \to \mathcal{Y}$, then $\eta_f(P_{g(X)|X}) = \eta_{\mathsf{TV}}(P_{g(X)|X}) = 1$ (cf., e.g., Proposition II.4.12 from [9]).

Generalization bound for noisy DNNs. According to the above, if all the feature maps $g_{\mathbf{w}_l}$, for $l = 1, \ldots, L$, in the DNN are deterministic, the contraction coefficients we are looking for degenerate to 1, landing us back at the bound from Theorem 1. To arrive at nontrivial contraction we consider a noisy DNN model where the feature map at each layer is perturbed by isotropic Gaussian noise, i.e.,

$$\widetilde{T}_l = T_l + \epsilon_l Z_l, \quad l = 1, \ldots, L, \tag{3}$$

where $Z_l \sim N(0, \mathbf{I}_{d_l})$ is independent of the input and $\epsilon_l \in \mathbb{R}_+$ is a constant. As before, we introduce the subscript i, writing $\widetilde{T}_{l,i}$, when the input to the DNN is X_i. Such noisy DNNs were explored in [16] and were shown to serve as good approximations of classical (deterministic) networks.

To analyze generalization error under the noisy DNN model, we present the following lemma that bounds the SDPI coefficient for the aforementioned channel.

Lemma 2 (SDPI coefficient bound)**.** *Let $X \sim P_X \in \mathcal{P}(\mathbb{R}^{d_x})$ and consider a bounded function $g : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$. Set $Y = g(X) + \epsilon N$, where $\epsilon > 0$ and $N \sim \mathcal{N}(0, \mathbf{I}_{d_y})$ is independent of X. The SDPI coefficient of the induced channel $P_{Y|X}$ satisfies*

$$\eta_f(P_{Y|X}) \le \eta_{\mathsf{TV}}(P_{Y|X}) \le 1 - 2\mathsf{Q}\left(\frac{\sqrt{2d_y}\|g\|_\infty}{2\epsilon}\right),$$

where $\mathsf{Q}(x) \coloneqq \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t$ is the Gaussian complimentary cumulative distribution function.

Lemma 2, which is proven in Appendix C, implies that whenever $\sqrt{d_y}\|g\|_\infty \epsilon^{-1} > 0$, we have $\eta_{\mathsf{KL}}(P_{Y|X}) < 1$. Consequently, the layer mappings in a noisy DNN with bounded activations present non-trivial contraction, which gives rise to the following result.

Theorem 3 (Noisy DNN generalization bound)**.** *Consider the noisy DNN model from* (3) *with bounded activation functions $\phi_l, l = 1, \ldots, L$. If 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*

$$|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \le \frac{\sigma\sqrt{2}}{n} \sum_{i=1}^n \sqrt{\prod_{l=1}^L \left(1 - 2\mathsf{Q}\left(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\right)\right) \mathsf{I}(X_i; \mathbf{W}|Y_i) + \mathsf{I}(Y_i; \mathbf{W})}.$$

The proof of Theorem 3 is provided in Appendix D. The argument initiates at the bound $\mathsf{UB}(L)$ from Theorem 1 and first factors out the terms that depend on the label Y from the KL divergence. This yields the summand $\mathsf{I}(Y_i; \mathbf{W})$. This step is necessary since the label is not processed by the noisy DNN, and the corresponding SDPI coefficient without factoring it out would trivialize to 1. For the remaining term, we invoke the SDPI L times collecting the coefficients and invoking Lemma 2 to arrive at the desired bound. We also note that $\|\phi_l\|_\infty$ is typically small, e.g., $\|\phi_l\|_\infty = 1$ if $\phi_l \in \{\mathrm{sigmoid}, \mathrm{softmax}, \tanh\}$.

Remark 2 (Interpretation and discussion)**.** *We make the following observations regarding the bound from Theorem 3. For fixed noise parameters $\epsilon_1, \ldots, \epsilon_L$, the coefficient decreases from 1 to 0 as the layer dimensions d_l shrink and the number of layers L grows. Similarly, since the label follows a categorical distribution of parameter K, we have $\mathsf{I}(Y_i; \mathbf{W}) \le \log K$. The smaller the number of distinct labels K, the better the generalization bound. The behavior of $\mathsf{I}(X_i; \mathbf{W}|Y_i)$, on the other hand, is harder to pin down as it depends on the data distribution and the learning algorithm at hand. For instance, if the DNN parameter space is constrained to be finite, e.g., $\mathcal{W} = [B]^{d_1 \times d_0} \times \cdots \times [B]^{d_L \times d_{L-1}}$ for some $B \in \mathbb{Z}_+$, then we can upper bound the mutual information by $\mathsf{I}(X_i; \mathbf{W}|Y_i) \le \mathsf{H}(\mathbf{W}) \le \sum_{l=1}^L d_l d_{l-1} \log B$, which increases as d_l and L grow. Consider increasing the depth of an L-layer neural net with discrete parameters as above, by adding an extra hidden layer with dimension d_{l^*}. Figure 2 plots the said mutual information bound and shows that if d_{l^*} is sufficiently small, then the generalization bound shrinks as a result of the added layer. It suggests that a deep but narrower network may generalize better. However, to draw more compelling conclusions, it is necessary to conduct thorough analyses and experiments for general algorithms/architectures. It may also be insightful to explore generalization bounds based on other divergences and to account for more activation functions.*

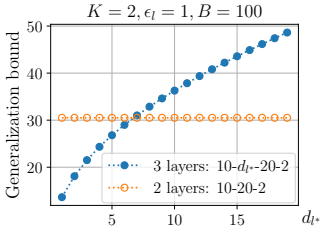

Figure 2: Generalization bounds w/wo an added layer for finite $\mathcal{W}$.

References

[1] Gholamali Aminian, Saeed Masiha, Laura Toni, and Miguel RD Rodrigues. Learning algorithm generalization error bounds via auxiliary distributions. *arXiv preprint arXiv:2210.00483*, 2022.

[2] Gholamali Aminian, Laura Toni, and Miguel RD Rodrigues. Information-theoretic bounds on the moments of the generalization error of learning algorithms. In *IEEE International Symposium on Information Theory (ISIT)*, 2021.

[3] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[4] Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, pages 7234–7243, 2018.

[5] SS Barsov and Vladimir V Ul'yanov. Estimates of the proximity of gaussian measures. In *Sov. Math., Dokl*, volume 34, pages 462–466, 1987.

[6] Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020.

[7] Satrajit Chatterjee and Piotr Zielinski. On the generalization mystery in deep learning. *arXiv preprint arXiv:2203.10036*, 2022.

[8] Eugenio Clerico, Amitis Shidani, George Deligiannidis, and Arnaud Doucet. Chained generalisation bounds. In *Conference on Learning Theory*, pages 4212–4257. PMLR, 2022.

[9] Joel Cohen, Johannes HB Kempermann, and Gheorghe Zbaganu. *Comparisons of stochastic matrices with applications in information theory, statistics, economics and population*. Springer Science & Business Media, 1998.

[10] Joel E. Cohen, Yoh Iwasa, Gh. Rautu, Mary Beth Ruskai, Eugene Seneta, and Gh. Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear Algebra and its Applications*, 179:211–235, 1993.

[11] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[12] Roland L Dobrushin. Central limit theorem for nonstationary markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

[13] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[14] Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 2021.

[15] Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

[16] Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *International Conference on Machine Learning*, pages 2299–2308. PMLR, 2019.

[17] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[18] Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33, 2020.

[19] Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 2020.

[20] Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems*, 34:24670–24682, 2021.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[22] Daniel Jakubovitz, Raja Giryes, and Miguel RD Rodrigues. Generalization error in deep learning. In *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*, pages 153–193. Springer, 2019.

[23] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. In *Mathematical Aspects of Deep Learning*. Cambridge University Press, 2022.

[24] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[25] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd international conference on artificial intelligence and statistics*, pages 888–896. PMLR, 2019.

[26] David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

[27] David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

[28] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[29] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

[30] Yury Polyanskiy and Yihong Wu. Dissipation of information in channels with input constraints. *IEEE Transactions on Information Theory*, 62(1):35–55, 2015.

[31] Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pages 211–249. Springer, 2017.

[32] Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

[33] Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

[34] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[35] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[37] Hao Wang, Mario Diaz, José Cândido S Santos Filho, and Flavio P Calmon. An information-theoretic view of generalization via Wasserstein distance. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 577–581. IEEE, 2019.

[38] Lei Wu, Zhanxing Zhu, and Weinan E. Towards understanding generalization of deep learning: Perspective of loss landscapes. *ICML 2017 Workshop on Principled Approaches to Deep Learning, Sydney, Australia*, 2017.

[39] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

[40] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[41] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. In *International Conference on Learning Representations*, 2018.

A Proof of Theorem 1

Let us rewrite the risks and generalization error under the DNN setup. Let $(X, Y) \sim P_{X,Y}$ be a pair of test data sample. At each layer l, the internal representation T_l of a test data feature X is conditionally independent of W_{l+1}^L given W_1^l. For any $\mathbf{W} \in \mathcal{W}$, let the loss function be rewritten as $\ell(\mathbf{W}, X, Y) = \ell(g_{\mathbf{W}_L} \circ g_{\mathbf{W}_{L-1}} \circ \cdots \circ g_{\mathbf{W}_1}(X), Y)$. The expected population risk over all possible $\mathbf{W}$ is given by

$$\begin{aligned}\mathbb{E}_W[\mathcal{L}_{\mathsf{P}}(\mathbf{W}, P_{X,Y})] &= \mathbb{E}[\ell(g_{\mathbf{W}_L} \circ g_{\mathbf{W}_{L-1}} \circ \cdots \circ g_{\mathbf{W}_1}(X), Y)] \\ &= \mathbb{E}[\ell(g_{\mathbf{W}_L} \circ g_{\mathbf{W}_{L-1}} \circ \cdots \circ g_{\mathbf{W}_{l+1}}(T_l), Y)] \\ &= \mathbb{E}[\mathbb{E}[\ell(g_{\mathbf{W}_L} \circ g_{\mathbf{W}_{L-1}} \circ \cdots \circ g_{\mathbf{W}_{l+1}}(T_l), Y)|\mathbf{W}_1^l]]\end{aligned}$$

where $l \in [L]$ and given $\mathbf{W}_1^l$, (T_l, Y) are independent of $\mathbf{W}_{l+1}^L$.

Denote the overall feature mapping function as $f_{\mathbf{W}} \triangleq g_{\mathbf{W}_L} \circ g_{\mathbf{W}_{L-1}} \circ \cdots \circ g_{\mathbf{W}_1}$. Similarly, for any $l \in [L]$, the expected empirical risk can also be rewritten as

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{\mathsf{E}}(\mathbf{W}, D_n)] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \ell(f_{\mathbf{W}}(X_i), Y_i)\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(f_{\mathbf{W}}(X_i), Y_i)] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(g_{\mathbf{W}_L} \circ g_{\mathbf{W}_{L-1}} \circ \cdots \circ g_{\mathbf{W}_{l+1}}(T_{l,i}), Y_i)] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\mathbb{E}[\ell(g_{\mathbf{W}_L} \circ g_{\mathbf{W}_{L-1}} \circ \cdots \circ g_{\mathbf{W}_{l+1}}(T_{l,i}), Y_i)|\mathbf{W}_1^l]\big].\end{aligned}$$

For notational simplicity, let $g_{\mathbf{W}_k^j} := g_{\mathbf{W}_k} \circ g_{\mathbf{W}_{k-1}} \circ \cdots \circ g_{\mathbf{W}_j}$ for any $k < j$ and $k, j \in \mathbb{N}$. Then the expected generalization error can be rewritten as

$$\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\mathbb{E}[\ell(g_{\mathbf{W}_{l+1}^L}(T_l), Y)|\mathbf{W}_1^l] - \mathbb{E}[\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}), Y_i)|\mathbf{W}_1^l]\right]. \quad (4)$$

If the loss function $\ell(\mathbf{w}, X, Y)$ is σ-sub-Gaussian under $P_{X,Y}$ for all $\mathbf{w} \in \mathcal{W}$, we also have for any $l \in [0 : L]$, $\ell(g_{\mathbf{w}_{l+1}^L}(T_l), Y)$ is σ-sub-Gaussian under $P_{T_l,Y|\mathbf{W}=\mathbf{w}}$ for all $\mathbf{w} \in \mathcal{W}$. From Donsker-Varadhan representation, we have for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}&\mathsf{D}_{\mathsf{KL}}(P_{\mathbf{W}_{l+1}^L, T_{l,i}, Y_i|\mathbf{W}_1^l} \| P_{T_l,Y|\mathbf{W}_1^l} \otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}) \\ &\geq \mathbb{E}_{\mathbf{W}_{l+1}^L, T_{l,i}, Y_i|\mathbf{W}_1^l}[\lambda\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}), Y_i)] - \log \mathbb{E}_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}\mathbb{E}_{T_l,Y|\mathbf{W}_1^l}[\exp(\lambda\ell(g_{\mathbf{W}_{l+1}^L}(T_l), Y))] \\ &\geq \lambda(\mathbb{E}_{\mathbf{W}_{l+1}^L, T_{l,i}, Y_i|\mathbf{W}_1^l}[\lambda\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}), Y_i)] - \mathbb{E}_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}\mathbb{E}_{T_l,Y|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_l), Y)]) - \frac{\lambda^2\sigma^2}{2}.\end{aligned}$$

We can decompose $\mathsf{D_{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})$ as follows

$$\begin{aligned}
&\mathsf{D_{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l}) \qquad (5)\\
&=\mathsf{D_{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_{l,i},Y_i|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})+\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\
&=\mathsf{I}(T_{l,i},Y_i;\mathbf{W}_{l+1}^L|\mathbf{W}_1^l)+\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l}).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathsf{I}(T_{l,i},Y_i;\mathbf{W}_{l+1}^L|\mathbf{W}_1^l)+\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\
&=\mathsf{D_{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\
&\geq\lambda\mathbb{E}_{\mathbf{W}_1^l}\left[\mathbb{E}_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}),Y_i)]-\mathbb{E}_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}\mathbb{E}_{T_l,Y|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_l),Y)]\right]-\frac{\lambda^2\sigma^2}{2}.
\end{aligned}$$

By optimizing the RHS over $\lambda>0$ and $\lambda\leq 0$, respectively, we finally obtain

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{W}_1^l}\mathbb{E}_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_{l,i}),Y_i)]-\mathbb{E}_{\mathbf{W}_1^l}\mathbb{E}_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}\mathbb{E}_{T_l,Y|\mathbf{W}_1^l}[\ell(g_{\mathbf{W}_{l+1}^L}(T_l),Y)]\right|\\
&\leq\sqrt{2\sigma^2\big(\mathsf{I}(T_{l,i},Y_i;\mathbf{W}_{l+1}^L|\mathbf{W}_1^l)+\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\big)},
\end{aligned}$$

which holds for all $l\in[L]$. Conditioned on $\mathbf{W}_l$, $T_{l,i}$ and T_l are generated by the same process from $T_{l-1,i}$ and T_{l-1}, respectively. By the data-processing inequality, the KL-divergence in (5) can be bounded as follows:

$$\begin{aligned}
&\mathsf{D_{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\
&\leq\mathsf{D_{KL}}(P_{\mathbf{W}_{l+1}^L,T_{l-1,i},Y_i|\mathbf{W}_1^l}\|P_{T_{l-1},Y|\mathbf{W}_1^l}\otimes P_{\mathbf{W}_{l+1}^L|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\\
&=\mathsf{D_{KL}}(P_{\mathbf{W}_l^L,T_{l-1,i},Y_i|\mathbf{W}_1^{l-1}}\|P_{T_{l-1},Y|\mathbf{W}_1^{l-1}}\otimes P_{\mathbf{W}_l^L|\mathbf{W}_1^{l-1}}|P_{\mathbf{W}_1^{l-1}})\\
&\vdots\\
&\leq\mathsf{D_{KL}}(P_{X_i,Y_i,\mathbf{W}_1^L}\|P_{X,Y}\otimes P_{\mathbf{W}_1^L})\\
&=\mathsf{I}(X_i,Y_i;\mathbf{W}).
\end{aligned}$$

Therefore, the expected generalization error in (4) can be upper bounded as follows:

$$\begin{aligned}
&|\mathsf{gen}(P_{\mathbf{W}|D_n},P_{X,Y})|\\
&\leq\frac{1}{n}\sum_{i=1}^n\sqrt{2\sigma^2\mathsf{D_{KL}}(P_{T_{L,i},Y_i|\mathbf{W}}\|P_{T_L,Y|\mathbf{W}}|P_{\mathbf{W}})}\\
&\leq\frac{1}{n}\sum_{i=1}^n\sqrt{2\sigma^2\big(\mathsf{I}(T_{L-1,i},Y_i;\mathbf{W}_L|\mathbf{W}_1^{L-1})+\mathsf{D_{KL}}(P_{T_{L-1,i},Y_i|\mathbf{W}_1^{L-1}}\|P_{T_{L-1},Y|\mathbf{W}_1^{L-1}}|P_{\mathbf{W}_1^{L-1}})\big)}\\
&\vdots\\
&\leq\frac{1}{n}\sum_{i=1}^n\sqrt{2\sigma^2\big(\mathsf{I}(T_{l,i},Y_i;W_l^L|\mathbf{W}_1^l)+\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})\big)}\\
&\vdots\\
&\leq\frac{1}{n}\sum_{i=1}^n\sqrt{2\sigma^2\big(\mathsf{I}(T_{1,i},Y_i;W_2^L|\mathbf{W}_1)+\mathsf{D_{KL}}(P_{T_{1,i},Y_i|\mathbf{W}_1}\|P_{T_1,Y|\mathbf{W}_1}|P_{\mathbf{W}_1})\big)}\\
&\leq\frac{1}{n}\sum_{i=1}^n\sqrt{2\sigma^2\mathsf{I}(X_i,Y_i;\mathbf{W})}.
\end{aligned}$$

B Proof for the Special Case of Discrete Latent Space

The information-theoretic quantities in $\mathsf{UB}(l)$ can be upper bounded as follows:

$$\mathsf{I}(T_{l,i},Y_i;\mathbf{W}_{l+1}^L|\mathbf{W}_1^l)+\mathsf{D_{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l}\|P_{T_l,Y|\mathbf{W}_1^l}|P_{\mathbf{W}_1^l})$$

$$
\begin{aligned}
&= \mathsf{I}(T_{l,i}; \mathbf{W}_{l+1}^L | \mathbf{W}_1^l) + \mathsf{I}(Y_i; \mathbf{W}_{l+1}^L | T_{l,i}, \mathbf{W}_1^l) + \mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l} \| P_{T_{l,i},Y|\mathbf{W}_1^l} | P_{\mathbf{W}_1^l}) \\
&= \mathsf{H}(T_{l,i}|\mathbf{W}_1^l) - \mathsf{H}(T_{l,i}|\mathbf{W}_1^l, \mathbf{W}_{l+1}^L) + \mathsf{H}(Y_i|T_{l,i}, \mathbf{W}_1^l) - \mathsf{H}(Y_i|T_{l,i}, \mathbf{W}_1^l, \mathbf{W}_{l+1}^L) \\
&\quad - \mathsf{H}(T_{l,i}, Y_i|\mathbf{W}_1^l) - \mathbb{E}_{P_{T_{l,i},Y_i,\mathbf{W}_1^l}}[\log P_{T_l,Y|\mathbf{W}_1^l}] \\
&= \mathsf{H}(T_{l,i}|\mathbf{W}_1^l) - \mathsf{H}(T_{l,i}|\mathbf{W}_1^l, \mathbf{W}_{l+1}^L) + \mathsf{H}(Y_i|T_{l,i}, \mathbf{W}_1^l) - \mathsf{H}(Y_i|T_{l,i}, \mathbf{W}_1^l, \mathbf{W}_{l+1}^L) \\
&\quad - \mathsf{H}(T_{l,i}|\mathbf{W}_1^l) + \mathsf{H}(Y_i|T_{l,i}, \mathbf{W}_1^l) - \mathbb{E}_{P_{T_{l,i},Y_i,\mathbf{W}_1^l}}[\log P_{T_l,Y|\mathbf{W}_1^l}] \\
&\leq 2\mathsf{H}(Y_i|T_{l,i}, \mathbf{W}_1^l) - \mathbb{E}_{P_{T_{l,i},Y_i,\mathbf{W}_1^l}}[\log P_{T_l,Y|\mathbf{W}_1^l}] \\
&\leq 2\log|\mathcal{Y}| - \mathbb{E}_{P_{T_{l,i},Y_i,\mathbf{W}_1^l}}[\log P_{T_l,Y|\mathbf{W}_1^l}] \\
&= 2\log K - \mathbb{E}_{P_{T_{l,i},Y_i,\mathbf{W}_1^l}}[\log P_{T_l,Y|\mathbf{W}_1^l}].
\end{aligned}
$$

Assuming $t_l(\mathbf{w}_1^l) \coloneqq \min_{t\in\mathcal{T}_l, y\in\mathcal{Y}} P_{T_l,Y|\mathbf{W}_1^l}(t, y|\mathbf{w}_1^l) \in \left(0, |\mathcal{T}_l \times \mathcal{Y}|^{-1}\right)$ and $\underline{t_l} \coloneqq \sup_{\mathbf{w}_1^l} t_l(\mathbf{w}_1^l)$, then

$$
\mathbb{E}_{P_{T_{l,i},Y_i,\mathbf{W}_1^l}}[\log P_{T_l,Y|\mathbf{W}_1^l}] \geq \log \underline{t_l}, \quad \text{and}
$$

$$
\mathsf{I}(T_{l,i}, Y_i; \mathbf{W}_{l+1}^L | \mathbf{W}_1^l) + \mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}_1^l} \| P_{T_l,Y|\mathbf{W}_1^l} | P_{\mathbf{W}_1^l}) \leq \log \frac{K^2}{\underline{t_l}}.
$$

C Proof of Lemma 2

When $Y = g(X) + \epsilon N$, where N is an independent noise and $\epsilon > 0$ is a constant controlling the signal-to-noise ratio (SNR), then $P_{Y|X} = P_{g(X)+\epsilon N}$. The Dobrushin's coefficient is given by

$$
\eta_{\mathsf{TV}}(P_{Y|X}) = \sup_{x,x'\in\mathbb{R}^{d_x}} \|P_{g(x)+\epsilon N} - P_{g(x')+\epsilon N}\|_{\mathsf{TV}}.
$$

Let N be a Gaussian noise generated from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$. Denote the vector function $g(\cdot)$ by $g(\cdot) = (g_1(\cdot), \ldots, g_{d_y}(\cdot))$. Following the proof in [30] and the total variation distance between two Gaussians with the same covariance matrix in [5, Theorem 1], we have for any $x, x' \in \mathcal{X}$,

$$
\begin{aligned}
\mathsf{D}_{\mathsf{TV}}(P_{g(x)+\epsilon N}, P_{g(x')+\epsilon N}) &= \|\mathcal{N}(g(x), \epsilon^2\mathbf{I}_{d_y}) - \mathcal{N}(g(x'), \epsilon^2\mathbf{I}_{d_y})\|_{\mathsf{TV}} \\
&= 1 - 2\mathsf{Q}\left(\frac{\|g(x) - g(x')\|}{2\epsilon}\right) \\
&\leq 1 - 2\mathsf{Q}\left(\frac{\sqrt{\sum_{i=1}^{d_y}(g_i(x) - g_i(x'))^2}}{2\epsilon}\right) \\
&\leq 1 - 2\mathsf{Q}\left(\frac{\sqrt{d_y(\|g(x)\|_\infty^2 + \|g(x')\|_\infty^2)}}{2\epsilon}\right) \\
&\leq 1 - 2\mathsf{Q}\left(\frac{\sqrt{2d_y}\|g\|_\infty}{2\epsilon}\right)
\end{aligned}
$$

where $\mathsf{Q}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\mathrm{d}t$ is the Gaussian complimentary CDF. Finally, we have

$$
\eta_f(P_{Y|X}) \leq \eta_{\mathsf{TV}}(P_{Y|X}) \leq 1 - 2\mathsf{Q}\left(\frac{\sqrt{2d_y}\|g\|_\infty}{2\epsilon}\right).
$$

D Proof of Theorem 3

Among the commonly used activation functions, the following functions and their gradients are bounded: for any $u \in \mathbb{R}$,

- sigmoid function: $\mathrm{sigmoid}(u) = \frac{1}{1+e^{-u}} \in [0, 1]$, $\mathrm{sigmoid}'(u) = \frac{e^{-u}}{(1+e^{-u})^2} \in [0, 1]$.
- softmax function: $\mathrm{softmax}(u)_i = \frac{e^{u_i}}{\sum_j e^{u_j}} \in [0, 1]$, $\mathrm{softmax}'(u)_i = \mathrm{softmax}(u)_i(1 - \mathrm{softmax}(u)_i) \in [0, 1]$.
- tanh function: $\tanh u = \frac{e^u - e^{-u}}{e^u + e^{-u}} \in [-1, 1]$, $\tanh'(u) = 1 - \tanh^2 u \in [0, 1]$.

As shown in (3), that is, $\widetilde{T}_l = \phi_l(\mathbf{W}_l \widetilde{T}_{l-1}) + \epsilon_l Z_l$ and from Lemma 2, the Dobrushin's coefficient at the l^{th} layer is upper bounded by

$$\eta_{\mathsf{TV}}(P_{\widetilde{T}_l|\widetilde{T}_{l-1},\mathbf{W}}) \le 1 - 2\mathsf{Q}\bigg(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\bigg),$$

where $\|\phi_l\|_\infty = 1$ for $\phi_l \in \{\text{sigmoid}, \text{softmax}, \tanh\}$.

In the following proof, at no risk of confusion, we let $T_l = \widetilde{T}_l$ and $T_{l,i} = \widetilde{T}_{l,i}$, for simplicity. Conditioned on $\mathbf{W}$, $P_{T_{l-1,i}|Y_i,\mathbf{W}} \neq P_{T_{l-1}|Y,\mathbf{W}}$ but $P_{T_{l,i}|T_{l-1,i},\mathbf{W}} = P_{T_l|T_{l-1},W}$. Thus, we have

$$\begin{aligned}
&\mathsf{D}_{\mathsf{KL}}(P_{T_{l,i}|Y_i,\mathbf{W}}\|P_{T_l|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) \\
&= \mathsf{D}_{\mathsf{KL}}(P_{T_{l,i}|T_{l-1,i},\mathbf{W}} \circ P_{T_{l-1,i}|Y_i,\mathbf{W}}\|P_{T_l|T_{l-1},W} \circ P_{T_{l-1}|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) \\
&\le \bigg(1 - 2\mathsf{Q}\bigg(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\bigg)\bigg)\mathsf{D}_{\mathsf{KL}}(P_{T_{l-1,i}|Y_i,\mathbf{W}}\|P_{T_{l-1}|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}).
\end{aligned}$$

By induction, if we have L layers

$$\begin{aligned}
\mathsf{D}_{\mathsf{KL}}(P_{T_{L,i}|Y_i,\mathbf{W}}\|P_{T_L|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) &\le \bigg(1 - 2\mathsf{Q}\bigg(\frac{\sqrt{2d_L}\|\phi_L\|_\infty}{2\epsilon_L}\bigg)\bigg)\mathsf{D}_{\mathsf{KL}}(P_{T_{L-1,i}|Y_i,\mathbf{W}}\|P_{T_{L-1}|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) \\
&\le \prod_{l=L-1}^{L}\bigg(1 - 2\mathsf{Q}\bigg(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\bigg)\bigg)\mathsf{D}_{\mathsf{KL}}(P_{T_{L-2,i}|Y_i,\mathbf{W}}\|P_{T_{L-2}|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) \\
&\vdots \\
&\le \prod_{l=1}^{L}\bigg(1 - 2\mathsf{Q}\bigg(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\bigg)\bigg)\mathsf{D}_{\mathsf{KL}}(P_{T_{0,i}|Y_i,\mathbf{W}}\|P_{T_0|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) \\
&= \prod_{l=1}^{L}\bigg(1 - 2\mathsf{Q}\bigg(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\bigg)\bigg)\mathsf{I}(X_i;\mathbf{W}|Y_i). \qquad (6)
\end{aligned}$$

Recall the upper bound (2) in Theorem 1:

$$\begin{aligned}
|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| &\le \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2\mathsf{D}_{\mathsf{KL}}(P_{T_{L,i},Y_i|\mathbf{W}}\|P_{T_L,Y|\mathbf{W}}|P_{\mathbf{W}})} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2(\mathsf{D}_{\mathsf{KL}}(P_{T_{L,i}|Y_i,\mathbf{W}}\|P_{T_L|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) + \mathsf{I}(Y_i;\mathbf{W}))}, \qquad (7)
\end{aligned}$$

where (7) follows since for any $l \in [L]$,

$$\begin{aligned}
\mathsf{D}_{\mathsf{KL}}(P_{T_{l,i},Y_i|\mathbf{W}}\|P_{T_l,Y|\mathbf{W}}|P_{\mathbf{W}}) &= \mathsf{D}_{\mathsf{KL}}(P_{T_{l,i}|Y_i,\mathbf{W}}P_{Y_i|\mathbf{W}}\|P_{T_l|Y,\mathbf{W}}P_{Y|W}|P_{\mathbf{W}}) \\
&= \mathsf{D}_{\mathsf{KL}}(P_{T_{l,i}|Y_i,W}\|P_{T_l|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) + \mathsf{D}_{\mathsf{KL}}(P_{Y_i|\mathbf{W}}\|P_Y|P_{\mathbf{W}}) \\
&= \mathsf{D}_{\mathsf{KL}}(P_{T_{l,i}|Y_i,W}\|P_{T_l|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}}) + \mathsf{I}(Y_i;\mathbf{W}).
\end{aligned}$$

It can be observed that the data-processing inequality is only applied to $\mathsf{D}_{\mathsf{KL}}(P_{T_{l,i}|Y_i,W}\|P_{T_l|Y,\mathbf{W}}|P_{Y_i,\mathbf{W}})$ since Y_i is not processed. Furthermore, since $Y_i \in \mathcal{Y} = [K]$ is a discrete random variable, we have $\mathsf{I}(Y_i;\mathbf{W}) \le \mathsf{H}(Y_i) \le \log K$.

By combining (6) with (7), the expected generalization error is upper bounded by

$$|\mathsf{gen}(P_{\mathbf{W}|D_n}, P_{X,Y})| \le \frac{\sqrt{2\sigma^2}}{n}\sum_{i=1}^{n}\sqrt{\prod_{l=1}^{L}\bigg(1 - 2\mathsf{Q}\bigg(\frac{\sqrt{2d_l}\|\phi_l\|_\infty}{2\epsilon_l}\bigg)\bigg)\mathsf{I}(X_i;\mathbf{W}|Y_i) + \log K}.$$

The proof is thus completed.