How Does Information Bottleneck Help Deep Learning?

Kenji Kawaguchi [*1] **Zhun Deng** [*2] **Xu Ji** [*3] **Jiaoyang Huang** [4]

Abstract

Numerous deep learning algorithms have been inspired by and understood via the notion of information bottleneck, where unnecessary information is (often implicitly) minimized while task-relevant information is maximized. However, a rigorous argument for justifying why it is desirable to control information bottlenecks has been elusive. In this paper, we provide the first rigorous learning theory for justifying the benefit of information bottleneck in deep learning by mathematically relating information bottleneck to generalization errors. Our theory proves that controlling information bottleneck is one way to control generalization errors in deep learning, although it is not the only or necessary way. We investigate the merit of our new mathematical findings with experiments across a range of architectures and learning settings. In many cases, generalization errors are shown to correlate with the degree of information bottleneck: i.e., the amount of the unnecessary information at hidden layers. This paper provides a theoretical foundation for current and future methods through the lens of information bottleneck. Our new generalization bounds scale with the degree of information bottleneck, unlike the previous bounds that scale with the number of parameters, VC dimension, Rademacher complexity, stability or robustness. Our code is publicly available at: https://github.com/xu-ji/information-bottleneck

1. Introduction

The information bottleneck principle (Tishby et al., 1999; Slonim & Tishby, 2000) has been a great concept in balancing the trade-off between the complexity of representation and the power of predicting. It is based on the notion of minimal sufficient statistics for extracting information about target $Y \in \mathcal{Y}$ into representation $Z = \phi(X) \in \mathcal{Z}$ from input $X \in \mathcal{X}$. An information bottleneck imposes regularization at representation Z by minimizing the mutual information between X and Z, $I(X;Z)$, while maximizing the mutual information between Y and Z, $I(Y;Z)$.

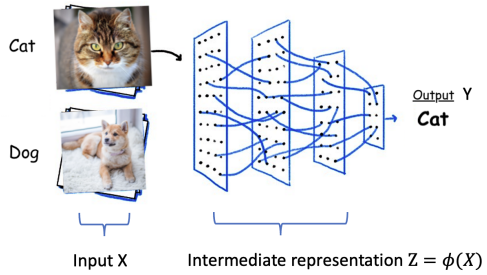


Figure 1. Illustration of X, Y and Z. This paper studies the relationship between performances of deep neural networks and the mutual information between X and Z. Our theory proves that controlling this mutual information is one way to control performances in deep learning, although it is not the necessary way.

In practice $I(X;Z)$ is often minimized implicitly, e.g. as a result of stochastic gradient descent (SGD) or an architecture choice (Shwartz-Ziv & Tishby, 2017). An explicit minimization of $I(X;Z)$ has been also adopted in the machine learning literature as a regularization technique (Alemi et al., 2016; 2018), where the mutual information is either estimated by averaging log probabilities of latent representations over empirical samples or replaced by a tractable upper bound (Kirsch et al., 2020; Kolchinsky & Tracey, 2017; Alemi et al., 2016). More generally, the notion of bottlenecks on representation expressivity has been used in work on structural inductive biases (Goyal & Bengio, 2022).

Consequently, understanding the connection between the information bottleneck regularizer $I(X;Z)$ and the generalization ability of machine learning models has become an active area of research. Given its importance, Shwartz-Ziv et al. (2019) provided the following conjecture:

Conjecture 1. ***(Informal version** (Shwartz-Ziv et al., 2019))* *With probability at least $1-\delta$ over the training data $s = \{(x_i, y_i)\}_{i=1}^n$ drawn from the same distribution as a random variable pair (X, Y), for the generalization error*

[*]Equal contribution. Author ordering determined by coin flip. [1]NUS [2]Columbia University [3]Mila [4]University of Pennsylvania. Correspondence to: Kenji Kawaguchi <kenji@nus.edu.sg>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

$\Delta(s) = \mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i)$*, there is a bound obeys the following form:*

$$\Delta(s) \leq \sqrt{\frac{2^{I(X;Z_l^s)} + \log\frac{2}{\delta}}{2n}}, \tag{1}$$

where f^s is the full model obtained by training and $Z_l^s = \phi_l^s(X)$ is the output of the an intermediate l-layer encoder ϕ_l^s of the model, i.e. representation obtained after passing through the first l layers.

However, this appealing conjecture cannot be applied to explain the success of information bottleneck principle in practice. First, the proof of the bound in this conjecture is *incomplete*. More importantly, as pointed out by Hafez-Kolahi et al. (2020), there is a critical drawback in the formulation of this conjecture: Shwartz-Ziv et al. (2019) implicitly assumes the independence of Z_l^s and s in the arguments of this conjecture, which means that they treated the encoder ϕ_l^s as fixed and independent of training data s. Indeed, Hafez-Kolahi et al. (2020) constructed a counterexample to show that the conjecture is invalid when the encoder ϕ_l^s is also learned with the training data s. This is because minimizing $I(X; Z_l^s)$ only does not sufficiently constrain the complexity of ϕ_l^s, allowing it to arbitrarily overfit to the training data with a large generalization gap, in contradiction to the inequality (1). In other words, when selecting the encoder's parameters is part of the learning problem, measuring compression via $I(X; Z_l^s)$ does not capture the degree of overfitting of the encoder's parameters.

Accordingly, as a first step towards proving a sample complexity bound via information bottleneck, Hafez-Kolahi et al. (2020) focused on the *input* layer and proved the following input compression bound for binary classification: if $\mathcal{Y} = \{0, 1\}$ and ℓ is the 0–1 loss, then for any $\delta > 0$, with probability at least $1 - \delta$ over the training dataset s, they roughly prove

$$\Delta(s) = \tilde{\mathcal{O}}\left(\sqrt{\frac{2^{H(X)}}{n}}\right), \tag{2}$$

up to a factor $2^{1/\epsilon}$ for some constant ϵ satisfying $\epsilon = \Omega\big(\sqrt{(2^{6H(X)/\epsilon} + \log(1/\delta) + 2)/n}\big)$. However, despite the popularity of the of information bottleneck principle and active usage in practice, this is still far from being a valid sample complexity bound, as noted by Hafez-Kolahi et al. (2020).

To the best of our knowledge, in the current literature, much of the work on information bottleneck assumes its benefits, but no rigorous and valid sample complexity bounds have been proposed to justify why it is desirable to control information bottlenecks. In this paper, we make the first step to fill in this gap and provide an answer to the following open problem:

"*How does information bottleneck help deep learning from the perspective of statistical learning theory?*"

As our ***first contribution***, we resolve this open question by providing novel and complete proofs for end-to-end learning of intermediate representations (Theorem 2). To the best of our knowledge, we provide the first rigorous generalization bound for information bottleneck in the case of learning representations, showing that simplicity in both the representation and representation function are factors that support generalization.

As our ***second contribution***, an intermediate step and byproduct of our novel proof for Theorem 2 not only completes the proof of Conjecture 1, where Z is treated as fixed random variable and independent of the training data s, it also significantly improve the previous bound in the conjecture. We show the generalization error roughly (with high probability) as

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{I(X; Z_l^s|Y) + 1}{n}}\right) \quad \text{as } n \to \infty.$$

This not only improves the numerator of the bound from an exponential dependence to linear dependence on mutual information, but also improves $I(X; Z_l^s)$ to a smaller quantity $I(X; Z_l^s|Y)$. More importantly, it is of independent interest and applicable to cases in transfer learning and unsupervised learning.

Finally, in Section 5, we consolidate our theoretical findings by comprehensive experiments on our bounds and related generalization prediction metrics, finding that empirical estimates of the main factors in our bounds are strong predictors of the generalization gap.

2. Preliminaries

In this section, we describe the notations we use and settings we mainly consider.

Notation. We are given a training dataset $s = ((x_i, y_i))_{i=1}^n \sim \mathcal{P}^{\otimes n}$ of n samples where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ are i.i.d. drawn from a joint distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. We want to analyze the generalization gap, i.e., the gap between the expected loss and the training loss, which is defined as

$$\Delta(s) := \mathbb{E}_{(X,Y)\sim\mathcal{P}}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i),$$

where $\ell : \mathbb{R}^{m_y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a bounded per-sample loss, and $f^s : \mathcal{X} \mapsto \mathbb{R}^{m_y}$ represents a deep neural network learned with a given training dataset s. Here, X and Y are the corresponding random variables for x_i and y_i with $(X, Y) \sim \mathcal{P}$. We use symbol $\circ$ to represent the composition

of functions and the notation of $[D+1] = \{1, 2, \ldots, D+1\}$. We define the random variable of the output of the l-th layer by

$$Z_l^s = \phi_l^s(X), \tag{3}$$

where ϕ_l^s is the map for the first l layer with with $\phi_l^s(x) \in \mathcal{Z}_l^s$. That is, for any layer index $l \in [D+1]$, we can decompose the neural network f^s by

$$f^s = g_l^s \circ \phi_l^s, \tag{4}$$

where g_l^s is the map for the rest of the layers after l layers. For convenience, we refer to ϕ_l^s as the encoder and to g_l^s as the decoder, though it is unnecessary to have an explicit structure of an encoder and a decoder. Here, the case of $l = 1$ corresponds to the input layer where $\phi_1^s(x) = x$ and $g_1^s(x) = f^s(x)$. The case of $l = D+1$ corresponds to the output layer where $\phi_{D+1}^s(x) = f^s(x)$ and $g_{D+1}^s(q) = q$. f^s can also be decomposed as

$$f^s = h_{D+1}^s \circ h_D^s \circ h_{D-1}^s \circ \cdots \circ h_1^s,$$

where h_l^s represents the computation of the l-th layer; i.e., $\phi_l^s = h_l^s \circ h_{l-1}^s \circ \cdots \circ h_1^s$ and $g_l^s = h_{D+1}^s \circ h_D^s \circ \cdots \circ h_{l+1}^s$.

We use $\mathcal{A}$ to denote the learning algorithm that returns the output functions of each layer; i.e., $\mathcal{A}(s) = \{h_l^s\}_{l=1}^{D+1}$. Then, by taking a subset of the output coordinates, we define $\tilde{\mathcal{A}}_l(s) = \{h_k^s\}_{k=1}^l$. Finally, by composing the outputs of $\tilde{\mathcal{A}}_l$, we define $\mathcal{A}_l(s) = h_l^s \circ h_{l-1}^s \circ \cdots \circ h_1^s = \phi_l^s \in \mathcal{M}_l$ (where $\{\mathcal{M}_l\}_l$'s are families of functions). Define the maximum loss

$$\mathcal{R}(f^s) = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \ell(f^s(x), y.)$$

We then define the random variable of the encoder of the l-th layer by

$$\phi_l^s = \mathcal{A}_l(s). \tag{5}$$

Presumption. Following the previous work (Shwartz-Ziv et al., 2019), we consider the setting of $|\mathcal{X}| < \infty$ and $|\mathcal{M}_l| < \infty$. This is the natural setting with digital computers (e.g., using floating point). In this setting, the mutual information that we consider are all finite and thus all bounds are nontrivial. Similar restrictions are commonly considered in theory work involving mutual information to avoid the issue of infinite mutual information; for example, in Xu & Raginsky (2017), they consider countable hypothesis space. We follow the above setting in our main results, but we also show that those requirements can be relaxed: see more details in Section 4 and Appendix E.2.

3. Main Results

In this section, we establish sample complexity bounds to connect information bottlenecks and generalization errors. We start with completing and improving the previous results in the setting where the encoder ϕ_l^s is treated as fixed and independent of training data s (Shwartz-Ziv et al., 2019; Hafez-Kolahi et al., 2020) in Section 3.1. It will serve as an important intermediate step toward our final result, where we extend the argument to deal with the main case of our interest – learning the encoder ϕ_l^s with s in Section 3.2.

3.1. Encoder independent with the training data

The following theorem shows that we can indeed optimize the control of expected loss by minimizing the conditional mutual information $I(X; Z_l^s|Y)$ and the training loss if the encoder ϕ_l^s is fixed and independent of training data s.[1] Even though this simplified case is just an intermediate step towards our final results, Theorem 1 is still useful and of independent interest. For example, it is applicable when the encoder is learned with data independent of s, such as in certain cases in transfer learning and unsupervised learning.

Theorem 1. *Let $l \in \{1, \ldots, D\}$. Suppose that ϕ_l^s is fixed independently of the training dataset s. Then, for any $\delta > 0$, with probability at least $1-\delta$ over training data s, the following holds:*

$$\Delta(s) \le G_3^l \sqrt{\frac{I(X; Z_l^s|Y)\ln(2) + \mathcal{G}_2^l}{n}} + \frac{G_1^l(0)}{\sqrt{n}}, \tag{6}$$

where $G_1^l(0) = \tilde{\mathcal{O}}(1)$, $\mathcal{G}_2^l = \tilde{\mathcal{O}}(1)$, and $G_3^l = \tilde{\mathcal{O}}(1)$, as $n \to \infty$. The formula of $G_1^l(0)$, $\mathcal{G}_2^l$ and G_3^l are given in Appendix E.1.

Theorem 1 rigorously completes the proof of Conjecture 1, with the significant improvements, which we will provide more detailed explanation in the following paragraph.

Explanation of Theorem 1. There are two significant improvements in our bound (6) when compared with the previous bound (1). First, we reduce the exponential dependence to the linear dependence by replacing the exponential growth rate $2^{I(X;Z_l^s)}$ with the linear growth rate $I(X; Z_l^s)$. Second, we replace $I(X; Z_l^s)$ with $I(X; Z_l^s|Y)$, which is the expected mutual information between X and Z_l^s conditioned on Y. To see why it is an improvement, notice that $I(X; Z_l^s|Y) \le I(X; Z_l^s)$ since we can decompose $I(X; Z_l^s)$ into two components by using the chain rule as in Federici et al. (2020): $I(X; Z_l^s) = I(X; Z_l^s|Y) + I(Y; Z_l^s)$. Here, $I(X; Z_l^s|Y) \ge 0$ is the superfluous information that we want to minimize so as to maximize the predictive information $I(Y; Z_l^s) \ge 0$. Therefore, the spirit of information bottleneck to regularize $I(X; Z_l^s)$ while maximizing $I(Y; Z_l^s)$ is an indirect way to regularize $I(X; Z_l^s|Y)$. Accordingly, instead of regularizing $I(X; Z_l^s)$, recent works

[1] Here, we still use the superscript s for various quantities to maintain notation consistency. Theorem 1 considers encoders that are independent of s, while Theorem 2 and the rest of this paper consider encoders that are dependent of s.

have also start to consider regularizing $I(X;Z_l^s|Y)$ (Fischer, 2020; Federici et al., 2020; Lee et al., 2021). In terms of theory, replacing $I(X;Z_l^s)$ with $I(X;Z_l^s|Y)$ is ***qualitatively significant*** because $I(X;Z_l^s)$ cannot be zero while maintaining the label-relevant information $I(Y;Z_l^s)$, unlike $I(X;Z_l^s|Y)$.

For practical use of our bound, one can think about the case of fixed-feature learning, where the representation is learnt by other dataset independent of s. This is widely used in transfer learning and pre-training, where a l-layer representation is learnt from a large public available dataset (e.g. ImageNet) and we further use task specific and possibly private dataset s to finetune extra few layers of the neural networks while fixing the representation.

3.2. Encoder learned with the training data

In the previous section, we have proven an improved version of Conjecture 1 for the setting with a fixed encoder ϕ_l^s. However, the typical usage of the information bottleneck principle is trying to minimizing $I(X;Z_l^s) = I(X;\phi_l^s(X))$ over the parameters of the encoder ϕ_l^s along with a discriminative objective. Thus, to support the typical usage of the information bottleneck principle, we need to extend the results to the setting of learning encoder ϕ_l^s with s. In this setting, the bound in Conjecture 1 is no longer valid as discussed in Section 1. However, the question about proving a sample complexity bound with the information bottleneck in the above typical setting is challenging and ***remains open***.

In this section, we present our main theorem that answers this open problem. Our result reconciles the information bottleneck regularizer $I(X;Z_l^s|Y)$ with the mutual information of the encoder and the training dataset $I(\phi_l^S;S)$.

Theorem 2 (**Main Theorem**). *Let $\mathcal{D} \subseteq \{1,2,\ldots,D+1\}$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the training set s, the following generalization bound holds:*

$$\Delta(s) \leq \min_{l\in\mathcal{D}} Q_l, \tag{7}$$

where for $l \leq D$,

$$Q_l = G_3^l \sqrt{\frac{\left(I(X;Z_l^s|Y)+I(\phi_l^S;S)\right)\ln(2)+\widehat{\mathcal{G}}_2^l}{n}} + \frac{G_1^l(\zeta)}{\sqrt{n}};$$

and for $l = D+1$,

$$Q_l = \mathcal{R}(f^s)\sqrt{\frac{I(\phi_l^S;S)\ln(2)+\check{\mathcal{G}}_2^l}{2n}},$$

Here, $S \sim \mathcal{P}^{\otimes n}$, $G_1^l(\zeta) = \tilde{\mathcal{O}}(\sqrt{I(\phi_l^S;S)+1})$, $\widehat{\mathcal{G}}_2^l = \tilde{\mathcal{O}}(1)$, $\check{\mathcal{G}}_2^l = \tilde{\mathcal{O}}(1)$, and $G_3^l = \tilde{\mathcal{O}}(1)$ as $n \to \infty$. The formulas of $G_1^l(\zeta)$, $\widehat{\mathcal{G}}_2^l$, $\check{\mathcal{G}}_2^l$, and G_3^l are given in Appendix E.1.

In Theorem 2, we denote by $S \sim \mathcal{P}^{\otimes n}$ the random variable following the same distribution as that of the training dataset $s \sim \mathcal{P}^{\otimes n}$. This notation is required here because the bound in Theorem 2 is equivalent to $\mathbb{P}_{s\sim\mathcal{P}^{\otimes n}}[\Delta(s) \leq g(I(X;Z_l^s|Y), I(\phi_l^S;S))] \geq 1-\delta$ for some function g. Here, the probability $\mathbb{P}_{s\sim\mathcal{P}^{\otimes n}}$ is taken with respect to s. The instantiations s within this $\mathbb{P}_{s\sim\mathcal{P}^{\otimes n}}$ are used in $\Delta(s)$ and $I(X;Z_l^s|Y)$, but not in $I(\phi_l^S;S)$ since $I(\phi_l^S;S)$ only depends the distribution $\mathcal{P}^{\otimes n}$ instead of the instantiations. That is, the reason why we need S here is similar to the same reason why we use j for the expression $\sum_{i=1}^n g_1(i, \sum_{j=1}^n g_2(j))$ (for some functions g_1, g_2) while we have $\sum_{j=1}^n g_2(j) = \sum_{i=1}^n g_2(i)$ in terms of its value, where i and j correspond to s and S, respectively.

In Theorem 2, the randomness of $I(X;Z_l^s|Y)$ and $I(\phi_l^S;S)$ are different. The mutual information $I(X;Z_l^s|Y)$ is calculated over the randomness of the distribution of the input X conditioning on Y, after fixing a realization of the training data s (for each fixed draw of s from $\mathbb{P}_{s\sim\mathcal{P}^{\otimes n}}$). In contrast, the mutual information $I(\phi_l^S;S)$ is computed over the randomness of the distribution of the training dataset $S \sim \mathcal{P}^{\otimes n}$.

Remark 1. One can consider the parameterization of the encoder as $\phi_l^S = \phi_{l,\theta_l^S}$ where θ_l^S is the parameter vector that is learned with S and contains all parameters of the layers up to l-th layer. In that case, Theorem 2 holds when replacing ϕ_l^S with θ_l^S.

Explanation of Theorem 2. Theorem 2 provides the ***first*** rigorous sample complexity bound for the information bottleneck in the setting of training the encoder ϕ_l^s with the same training data s. Here, Z_l^s is the random variable of the l-th layer's representation with dependence on the given training dataset s, and D is the number of all layers, including the input layer and excluding the output layer; i.e., Z_1^s is the input layer, Z_D^s is the last hidden layer, and Z_{D+1}^s is the output layer. Here, $I(\phi_l^S;S)$ is measuring the effect of overfitting the encoder, which is necessary to avoid the counter-example (Hafez-Kolahi et al., 2020, Example 3.1).

The main factor in the above theorem is $I(X;Z_l^s|Y) + I(\phi_l^S;S)$. This term captures the novel relationship that has not been studied in any previous sample complexity bounds. Specifically, this captures the relationship between "how much information from the input X the trained encoder ϕ_l^s retains, i.e., $I(X;Z_l^s|Y)$" and "how much information from the training dataset S is used to train the encoder ϕ_l^S, i.e., $I(\phi_l^S;S)$".

Theorem 2 is applicable when the encoder is trained with s and potentially additional data independent of s: e.g., supervised learning, semi-supervised learning, unsupervised learning, and transfer learning. For example, Theorem 2 captures the benefit of transfer learning in both terms of

$I(X; Z_l^s|Y)$ and $I(\phi_l^S; S)$ since the encoder ϕ_l^S is expected to have less dependence on S (target data) (for some $l \leq D$) in transfer learning, which tends to decrease $I(\phi_l^S; S)$.

Finally, we note that in the formula of $\widehat{\mathcal{G}}_2^l$, we have a linear dependence on $H(Z_l^s|X, Y)\ln(2)$ (see Appendix E.1). However, we have $H(Z_l^s|X, Y) = 0$ if the function ϕ_l^s is deterministic, which is the typical case for deep neural networks, because ϕ_l^s is the function used at inference or test time as opposed to training time (when dropout for example can be used). When the function ϕ_l^s is stochastic at test time, we have $H(Z_l^s|X, Y) \approx 0$ when the injected noise is small, and more generally $H(Z_l^s|X, Y) = \mathcal{O}(1)$ as $n \to \infty$.

4. Extensions

Our results thus far focus on the case of $|\mathcal{X}| < \infty$, which is already general enough to cover the realistic implementation on a computer and commonly considered in previous theory work (Shwartz-Ziv et al., 2019). Indeed, our presumption makes sure that the mutual information is finite and thus the bounds provided are non-trivial. In the case of $|\mathcal{X}| = \infty$, the mutual information can be infinite, and thus requires a separate treatment. In this section, we show how to generalize our arguments to the case of $|\mathcal{X}| = \infty$.

4.1. Neural networks with ReLU activation functions

First, we show that finite mutual information can be obtained in some cases even for the case of $|\mathcal{X}| = \infty$. Specifically, the following proposition shows that a (deterministic) neural network can have finite mutual information with ReLU activations with continuous distributions.

Proposition 1. *For a given neural network with ReLU activation functions, there are infinitely many continuous distributions over $\mathcal{X}$ such that the corresponding $I(X, Z|Y)$ is finite.*

4.2. Modification for valid bounds in the case of infinite mutual information

The mutual information for the information bottleneck is finite for many practical cases including the cases of discrete domains $\mathcal{X}$ with any models and of continuous domains $\mathcal{X}$ with stochastic models as well as the case in Proposition 1 with ReLU. However, it is infinite for some special case, for example, of continuous domains $\mathcal{X}$ with deterministic neural networks with certain types of injective activations such as sigmoid (instead of ReLU) (Amjad & Geiger, 2019). This subsection demonstrates that our bounds can be modified to produce finite bounds even for any special cases of the mutual information being infinite. Our results (Theorems 1–2 with Corollary 1) also resolve the known issue of arbitrariness of the mutual information with different binning methods (Saxe et al., 2019).

Consider an arbitrary (continuous or discrete) domain $\mathcal{X}$ and an arbitrary encoder $\tilde{\phi}_l^s$ such that $\tilde{\phi}_l^s(x) \in \tilde{\mathcal{Z}}_l^s$ and the set $\tilde{\mathcal{Z}}_l^s$ is potentially (uncountably or countably) infinite. Define the corresponding model $\tilde{f}^s$ by $\tilde{f}^s = g_l^s \circ \tilde{\phi}_l^s$ and $\tilde{Z}_l^s = \tilde{\phi}_l^s \circ X$. We formalize an arbitrary binning method $\mathcal{E}_l[\tilde{\phi}_l^s]$ of computing the mutual information (Chelombiev et al., 2019) as follows: for any $(l, \tilde{\phi}_l^s)$, let $\mathcal{E}_l[\tilde{\phi}_l^s] : \tilde{\mathcal{Z}}_l^s \to \mathcal{Z}_l^s \subseteq \tilde{\mathcal{Z}}_l^s$ be a function such that $|\mathcal{Z}_l^s| < \infty$. Set $\phi_l^s = \mathcal{E}_l[\tilde{\phi}_l^s] \circ \tilde{\phi}_l^s$; i.e., it follows that $Z_l^s = \mathcal{E}_l[\tilde{\phi}_l^s] \circ \tilde{Z}_l^s$ and $f^s = g_l^s \circ \mathcal{E}_l[\tilde{\phi}_l^s] \circ \tilde{\phi}_l^s$. Let $\hat{Q}_l$ and $\min_{l \in \mathcal{D}} Q_l$ be the right-hand side of Eq. (6) and Eq. (7) in Theorems 1–2 with this choice of encoder ϕ_l^s; i.e., $\hat{Q}_l$ and Q_l contain $I(X; Z_l^s|Y)$ instead of $I(X; \tilde{Z}_l^s|Y)$. Here, $I(X; Z_l^s|Y)$ is the mutual information computed by the binning method $\mathcal{E}_l[\tilde{\phi}_l^s]$ while $I(X; \tilde{Z}_l^s|Y)$ is the true mutual information of $\tilde{f}^s$. Let C_l be a nonnegative real number such that $\mathbb{P}(|\ell((g_l^s \circ \tilde{\phi}_l^s)(X), Y) - \ell((g_l^s \circ \mathcal{E}_l[\tilde{\phi}_l^s] \circ \tilde{\phi}_l^s)(X), Y)| \leq C_l) = 1$.

Corollary 1 shows that even when the mutual information $I(X; \tilde{Z}_l^s|Y)$ of the original model $\tilde{f}^s$ is infinite, Theorems 1–2 provide the finite bounds on the *original* model $\tilde{f}^s$ using the finite mutual information $I(X; Z_l^s|Y)$ returned by a binning method $\mathcal{E}_l[\tilde{\phi}_l^s]$:

Corollary 1. *Suppose that $C_l < \infty$. Then, Theorems 1–2 hold true also when we replace*

(Theorem 1) *Eq. (6) with $\Delta(s) \leq \hat{Q}_l + 2C_l < \infty$, and,*

(Theorem 2) *Eq. (7) with $\Delta(s) \leq \min_{l \in \mathcal{D}} Q_l + 2C_l < \infty$.*

The assumption on the finiteness of C_l is satisfied for common scenarios. For example, let L be the Lipschitz constant of the function $q \mapsto \ell(g_l^s(q), Y)$ w.r.t. some metric $d_{\mathcal{E}}$ almost surely (Fazlyab et al., 2019; Latorre et al., 2019; Aziznejad et al., 2020; Pauli et al., 2021). Set $\mathcal{E}_l[\tilde{\phi}_l^s]$ such that the radius of each bin w.r.t. the metric $d_{\mathcal{E}}$ is at most $\epsilon/\sqrt{nL^2}$ for some $\epsilon > 0$. We can then set $C_l = \epsilon/\sqrt{n}$.

In Corollary 1, the arbitrariness with binning methods is resolved: e.g., increasing the bin size ϵ can decrease the mutual information, but it also increases the value of $C_l = \epsilon/\sqrt{n}$. Thus, there is always a tradeoff and we can't arbitrarily change values of our bounds by choosing different binning methods. Similarly, for the case of infinite mutual information, we prove the validity of general methods of computing mutual information, including those of injecting noises and kernel density estimations in Appendix E.2.

5. Experiments

We conduct empirical experiments to investigate the following questions: (i) Does the information bottleneck regularizer $I(X; Z_l^s)$ alone reliably predict generalization when the encoder ϕ_l^s is learned with s? (ii) Does the main factor in our bound in Theorem 2 $\min_{l \in [D]} I(S; \theta_l^S) + I(X; Z_l^s|Y)$

Pearson Correlation	
Num. params.	-0.0294
$\prod_{\mathbf{l}} \|\theta^s_{\mathbf{l}}\|_F$	-0.0871
$\breve{I}(X; Z^s_l)$	0.3712
$\breve{I}(X; Z^s_l \| Y)$	0.3842
$\breve{I}(S; \theta^{\mathbf{S}}_{D+1})$	0.0091
$\breve{I}(S; \theta^S_l)$	0.0211
$\tilde{I}(S; \theta^S_l) + \breve{I}(X; Z^s_l)$	**0.3928**
$\tilde{I}(S; \theta^S_l) + \breve{I}(X; Z^s_l \| Y)$	**0.4130**

Table 1. Pearson correlation coefficient between metrics and the generalization gap in loss for constrained models trained for 5 class classification on 2D inputs. Positive values denote positive correlations. $\theta_{\mathbf{l}}$ denotes parameters of layer $\mathbf{l}$ and θ_l denotes parameters up to layer l.

with Remark 1 predict generalization more accurately than $I(X; Z^s_l)$ alone (or $I(X; Z^s_l|Y)$ alone)? (iii) How does varying layer l within the network affect the values of $I(S; \theta^S_l)$ and $I(X; Z^s_l)$ and their predictive ability?

5.1. On the Representation Compression Bound

As discussed in Sections 3, $I(X; Z^s_l)$ is generally not a reliable predictor of generalization because feature compression does not prevent the overfitting of the representation function's parameters. We investigate this further by designing a learning algorithm that trains models under various hyperparameter settings with the constraint that estimated $I(X; Z^s_l)$ is approximately constant. Following previous work such as Galloway et al. (2022), we use correlation analysis to empirically evaluate how strongly metrics predict generalization.

The inference problem studied was 5 class classification on clustered 2D inputs (Fig. 3). The model architecture was a 5 layer MLP with deterministic weights and feature layer l was fixed to the penultimate layer. Given training dataset s, each model q_{θ^s} was optimized with the cross-entropy loss $\min_{\theta^s} -\frac{1}{|s|}\sum_{(x,y)\in s}(\log(1/k)\sum_{j=1}^{k} q_{\theta^s}(y|z^j))$ s.t. $\hat{I}(X; Z^s_l) = \rho$, where features $z^j \sim q_{\theta^s}(Z^s_l|x)$, and $q_{\theta^s}(Z^s_l|x)$ is a multivariate Normal distribution with mean and variance computed by the MLP. Here,

$$\hat{I}(X; Z^s_l) = \frac{1}{|s|}\sum_{(x,y)\in s}\frac{1}{k}\sum_{j=1}^{k}\log\frac{q_{\theta^s}(z^j|x)}{\frac{1}{|s|}\sum_{(x',y')\in s} q_{\theta^s}(z^j|x')}$$

is a Monte-Carlo sampling based estimator of $I(X; Z^s_l)$, and constraint ρ was set to 1.5, approximately half the value of $\hat{I}(X; Z^s_l)$ attained without constraining $\hat{I}(X; Z^s_l)$. The neural network infers a distribution over a stochastic latent features so that $\hat{I}(X; Z^s_l)$ can be regularized and evaluated directly during training; in Section 5.2 we consider the case of deterministic features without regularization of $\hat{I}(X; Z^s_l)$. Whereas $\hat{I}(X; Z^s_l)$ is a sampling based estimator, we also use the upper-bound based estimator: $\breve{I}(X; Z^s_l) = \frac{1}{|s|}\sum_{(x,y)\in s}\frac{1}{k}\sum_{j=1}^{k}(\log q_{\theta^s}(z^j|x) - (\frac{1}{|s|}\sum_{(x',y')\in s}\log q_{\theta^s}(z^j|x')))$. We define $\hat{I}(X; Z^s_l|Y)$ and $\breve{I}(X; Z^s_l|Y)$ accordingly by conditioning these quantities on Y: $\hat{I}(X; Z^s_l|Y) = \frac{1}{|s|}\sum_{c\in C}\sum_{(x,y)\in s_c}\frac{1}{k}\sum_{j=1}^{k}\log\frac{q_{\theta^s}(z^j|x)}{\frac{1}{|s_c|}\sum_{(x',y')\in s_c} q_{\theta^s}(z^j|x')}$, and $\breve{I}(X; Z^s_l|Y) = \frac{1}{|s|}\sum_{c\in C}\sum_{(x,y)\in s_c}\frac{1}{k}\sum_{j=1}^{k}(\log q_{\theta^s}(z^j|x) - (\frac{1}{|s_c|}\sum_{(x',y')\in s_c}\log q_{\theta^s}(z^j|x')))$.

For the computation of $I(S; \theta^S_l)$, the learning algorithm is defined by the posterior distribution over network parameters $\mathbb{P}(\theta^S_l|S = s)$, which was modelled using SWAG (Maddox et al., 2019; Mandt et al., 2017), chosen for its popularity and simplicity. We denote the estimator of $I(S; \theta^S_l)$ using SWAG by $\breve{I}(S; \theta^S_l)$, where datasets S were drawn from the set of training datasets. To account for different scales of different estimation procedures, we tested rescaling $\breve{I}(S; \theta^S_l)$ by the average value of $\hat{I}(X; Z^s_l|Y)$, denoting rescaled values by $\tilde{I}(S; \theta^S_l)$ (see Appendix A for more details). 216 models were trained over varying architectures, weight decay rates, dataset draws, and random seeds. Model parameters were optimized end-to-end using the reparameterization trick (Kingma et al., 2015) with dual gradient descent for MI constraints (Bertsekas, 2014) (Appenix A). For each model, we measured the generalization gap between the test set and train set losses. We found that combining model compression and representation compression yielded the best predictor of generalization overall, and that this outperformed using representation compression alone (Table 1, 5). Additional experimental results on MNIST and Fashion MNIST datasets are given in Appendix B, showing that this conclusion also holds for stochastic feature networks in cases when $I(X; Z^s_l)$ is unconstrained.

5.2. Image Classification with DNNs

To investigate a common setting, we trained 540 deep neural networks on CIFAR10 without explicitly constraining MI, over varying preactivation ResNet architectures (He et al., 2016), weight decay rates, batch sizes, dataset draws and random seeds. To study representation compression by estimating MI with deterministically computed features, noise is customarily injected purely for analysis purposes (Saxe et al., 2019). We tested adaptive kernel density estimation (KDE) (Chelombiev et al., 2019), which models the latent represenation of an input as a unimodal Gaussian centred at the deterministic feature, with variance σ^2_l determined by scaling a base value according to maximum observed activation value in the layer. We also tested selecting σ^2_l by maximum likelihood estimation (MLE) of observed features under the constraint that estimated MI decreases with layer,

Generalization gap:	Spearman corr. Loss	Spearman corr. Error	Pearson corr. Loss	Pearson corr. Error	Kendall corr. Loss	Kendall corr. Error
$\frac{1}{D}\sum_{l=1}^{D} \breve{I}(X;Z_l^s)$	0.8481	0.7410	0.2116	0.1831	0.6425	0.5436
$\min_{l\in[D]} \breve{I}(X;Z_l^s)$	0.7145	0.5602	0.7203	0.5719	0.4461	0.3404
$\frac{1}{D}\sum_{l=1}^{D} \breve{I}(X;Z_l^s\|Y)$	0.8481	0.7406	0.2140	0.1853	0.6427	0.5435
$\min_{l\in[D]} \breve{I}(X;Z_l^s\|Y)$	0.7004	0.5434	0.7062	0.5560	0.4386	0.3305
$\breve{I}(S;\theta_{D+1}^S)$	0.4688	0.3112	0.2512	0.0775	0.2121	0.1208
$\min_{l\in[D]} \breve{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s\|Y)$	0.8434	0.7313	0.8437	0.7195	0.6270	0.5332
$\bar{I}(S;\theta_{D+1}^S)$	0.5370	0.3800	0.2924	0.1218	0.2442	0.1526
$\min_{l\in[D]} \bar{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s\|Y)$	**0.8632**	**0.7576**	**0.8511**	**0.7562**	**0.6626**	**0.5664**

Table 2. Correlation results across metrics for CIFAR10 models. Each value is in [-1, 1] and > 0 indicates positive correlation. Best metric highlighted. More results can be found in Appendix D.

Generalization gap:	Spearman corr. Loss	Spearman corr. Error	Pearson corr. Loss	Pearson corr. Error	Kendall corr. Loss	Kendall corr. Error
$\frac{1}{D}\sum_{l=1}^{D} \bar{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s\|Y)$	0.4429	0.2908	0.2783	0.1059	0.2349	0.1426
$\max_{l\in[D]} \bar{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s\|Y)$	0.5711	0.4204	0.2993	0.1311	0.2886	0.1945
$\min_{l\in[D]} \bar{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s\|Y)$	**0.8632**	**0.7576**	**0.8511**	**0.7562**	**0.6626**	**0.5664**
$\bar{I}(S;\theta_1^S) + \breve{I}(X;Z_1^s\|Y)$	0.6476	0.5292	0.1557	0.1331	0.4307	0.3504
$\bar{I}(S;\theta_D^S) + \breve{I}(X;Z_D^s\|Y)$	0.5711	0.4204	0.2993	0.1311	0.2886	0.1945

Table 3. Correlation results for $\bar{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s|Y)$ for CIFAR10 models across different layer summarization methods.

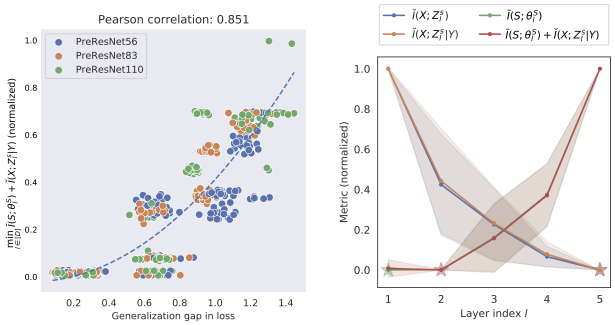


Figure 2. (**Left**) Results for $\min_{l\in[D]} \bar{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s|Y)$ for unconstrained models trained on CIFAR10. Dashed line denotes best polynomial fit with degree 2. (**Right**) Metrics averaged over models for each layer index. Star denotes minimum point for each metric. Values are normalized by subtracting the minimum and dividing by the range.

which follows from the information processing inequality. We report the results in this section for MLE and in Appendix D.4 for adaptive KDE. Representations were taken from $D = 5$ layers in the model, ranging from the input to the output of the penultimate layer. Again, SWAG was used to model the posterior $\mathbb{P}(\theta_l^S|S = s)$.

Since SWAG approximates the stationary distribution of SGD from a fixed initialization as a unimodal Gaussian (Mandt et al., 2017), we also tested averaging over initializations to obtain a richer posterior model, and denote the estimator of MI from this model as $\bar{I}(S;\theta_l^S)$, defined in Appendix D.2. To construct multiple training datasets, we sampled 5 training sets of size 15K from the CIFAR10 training set, and each test set was the original 10K test set.

The generalization gap was positively correlated with metrics measuring representation compression, but more correlated with metrics that combined both representation and model compression (Table 2). By increasing the value of layer index l of the encoder, MI between the encoder and training dataset increased, while MI between the representation and input decreased (Fig. 2), capturing a trade-off between these two measures of compression. For selection of hyperparameters σ_l^2, MLE (Fig. 2, Table 2, 3, 19, 17) outperformed adaptive KDE (Table 20, 18), however regardless of which scheme was used, the best metric that combined representation compression and model compression outperformed the best metric for representation compression or model compression individually. $\min_{l\in[D]} \bar{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s|Y)$ performed best overall (Fig. 2). Taking the minimum over layers (Theorem 2) outperformed other layer summarization methods (Table 3).

5.3. Feature binning experiments

To test whether the importance of model compression would hold when $I(X; Z_l^s)$ was estimated by binning deterministic features into discrete categories (Section 4.2), we conducted toy clustering experiments (Appendix 5.1) on deterministic feature models. The binning implementation of Saxe et al. (2019) was reused to discretize the activity of each node into 10 buckets and the information bottleneck was implemented using the surrogate objective of Kirsch et al. (2020, Eq. 1). 216 models were trained across MLP architectures and hyperparameters and SWAG was used to estimate the posterior. Model details and results are given in appendix C. We found that imposing the information bottleneck regularizer decreased the generalization prediction performance of feature compression metrics $\hat{I}(X; Z_l^s)$ and $\hat{I}(X; Z_l^s|Y)$, and combining model compression with feature compression metrics increased performance.

6. Proof Sketch

In this section, we further provide proof sketches of Theorems 1-2 for readers interested in having an overview of the proofs. The complete proofs are in Appendix F.

6.1. Proof Sketch of Theorem 1

Fix $l \in [D]$ and ϕ_l^s independently of the training dataset s. Let T be the standard *typical set* of Z_l^s in information theory. As we will see in the later sketch, we combine deterministic decompositions and probabilistic bounds with respect to the randomness of new fresh samples X and datasets s. The usages of probabilistic bounds for different sample spaces enable the exponential improvement over the previous bounds.

Step 1. Decompose the generalization gap into two terms as $\Delta(s) = A + B$, where A corresponds to the case of $Z_l^s \in T$, while B is for the case of $Z_l^s \notin T$. We will bound A and B separately.

Step 2. By standard argument from information theory, we have $\mathbb{P}(Z_l^s \notin T \mid Y = y) \leq \mathcal{O}(\frac{1}{\sqrt{n}})$ where the probability is with respect to X, with which we can roughly argue [2] that $B \leq \mathcal{O}(\frac{1}{\sqrt{n}})$.

Step 3. To bound A, we argue that A is bounded by a concentration gap of a special multinomial distribution over the elements of T, which is bounded roughly as $A = \mathcal{O}(\sqrt{(\ln |T|)/n})$ (with high probability with respect to S), by using a recent statistical result on multinomial distributions (Kawaguchi et al., 2022a, Lemma 3 & Proposition 3). Then, the standard argument from information theory approximately bounds the size of the typical set as $|T| \leq 2^{I(X;Z_l^s|Y)+C_T}$ for some $C_T > 0$, approximately resulting in: with high probability

$$A = \mathcal{O}\left(\sqrt{\frac{\ln |T|}{n}}\right) = \tilde{\mathcal{O}}\left(\sqrt{\frac{(I(X; Z_l^s|Y) + 1)}{n}}\right).$$

Step 4. Finally, By combining the above bounds on A and B, we approximately conclude the result.

[2] Although this requires a refinement of the standard argument. In Appendix F, we refine the argument using the McDiarmid's inequality and a further decomposition of B.

6.2. Proof Sketch of Theorem 2

Based on the result of Theorem 1 for fixed $l \in [D]$ and ϕ_l^s, we further generalize it for flexible l and learnable ϕ_l^s. Our proof is based on the tricky usage of probabilistic bounds for different sample spaces in Theorem 1's proof.

Step 1. Let $l \in [D]$. We first find a hypothesis space Φ_δ^l such that $\mathbb{P}(\phi_l^S \notin \Phi_\delta^l) \leq \delta$ and $|\Phi_\delta^l| \leq 2^{I(\phi_l^S;S)+C_\delta}$ for some $C_\delta \geq 0$. We then construct the corresponding hypothesis space $\mathcal{H}$ by $\mathcal{H} = \cup_{\phi_l \in \Phi_\delta^l} \mathcal{H}_{\phi_l}$ where $\mathcal{H}_{\phi_l} = \{g_l \circ \phi_l \mid g_l : \mathcal{Z}_l \to \mathbb{R}^{m_y}\}$.

Step 2. We then obtain the sample complexity bound for each $\mathcal{H}_{\phi_l}$ (for each $\phi_l \in \Phi_\delta^l$) by using the result of Theorem 1. For each $\phi_l \in \Phi_\delta^l$ that is fixed independently of s; i.e., $\mathbb{P}(\forall f \in \mathcal{H}_{\phi_l}, \mathcal{B}(f) \leq J_l(\delta)) \geq 1 - \delta$ where $\mathcal{B}(f) = \mathbb{E}_{X,Y}[\ell(f(X), Y)] - \frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)$ and $J_l(\delta)$ is the right-hand side of Eq. (6). Then, by taking union bound with the equal weighting over all $\phi_l \in \Phi_\delta^l$, we have $\mathbb{P}(\forall f \in \mathcal{H}, \mathcal{B}(f) \leq J_{\delta,l}) \geq 1 - \delta$, where $J_{\delta,l} = J_l(\delta/(2^{I(\phi_l^S;S)+C_\delta}))$.

Step 3. We now want to show that this bound holds for $\mathcal{B}(f^s)$ instead of $\mathcal{B}(f)$ for $f \in \mathcal{H}$. This is achieved if $f^s \in \mathcal{H}$. Since $\mathbb{P}(f^S \in \mathcal{H}) \geq 1 - \delta$ from the construction of $\mathcal{H}$ and $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$ for any events A and B, the following holds:

$$\begin{aligned}
&\mathbb{P}(\mathcal{B}(f^S) \leq J_{\delta,l}) \\
&\geq \mathbb{P}(f^S \in \mathcal{H})\mathbb{P}(\mathcal{B}(f^S) \leq J_{\delta,l} \mid f^S \in \mathcal{H}) \\
&\geq \mathbb{P}(f^S \in \mathcal{H})(1 - \delta) \geq 1 - 2\delta.
\end{aligned}$$

Therefore, by replacing δ with $\delta/2$, we have $\mathbb{P}(\mathcal{B}(f^S) \leq J_{\delta/2,l}) \geq 1 - \delta$.

Step 4. For the case of $l = D + 1$, the proof is significantly simplified because an entire model is an encoder as $f = \phi_{D+1}$; i.e., we replace the result of Theorem 1 with Hoeffding's inequality to conclude that $\mathbb{P}(\forall f \in \mathcal{H}_{\phi_{D+1}}, \mathcal{B}(f) \leq J_{D+1}(\delta)) \geq 1 - \delta$ where $J_{D+1}(\delta) = \mathcal{R}(f)\sqrt{(\ln(1/\delta))/(2n)}$. Using the same steps as the case of $l \in [D]$, we prove that $\mathbb{P}(\mathcal{B}(f^S) \leq J_{\delta/2,D+1}) \geq 1 - \delta$, where $J_{\delta,D+1} = J_{D+1}(\delta/(2^{I(\phi_l^S;S)+C_\delta}))$. By taking

union bounds over $l \in \mathcal{D} \subseteq \{1, 2, \dots, D+1\}$, we conclude $\mathbb{P}(\forall l \in \mathcal{D}, \mathcal{B}(f^S) \leq J_{\delta/(2|\mathcal{D}|),l}) = \mathbb{P}(\mathcal{B}(f^S) \leq \min_{l \in \mathcal{D}} J_{\delta/(2|\mathcal{D}|),l}) \geq 1 - \delta$.

Step 5. Organizing the expression of $J_{\delta/(2|\mathcal{D}|),l}$ yields the right-hand side of eq. (7), which proves Theorem 2.

7. Related Works

The implicit minimization of mutual information $I(X;Z)$ has been studied with the motivation of understanding why deep learning works through the lens of information bottlenecks (Shwartz-Ziv & Tishby, 2017). The previous work assumes the benefit of minimizing $I(X;Z)$ and questioned whether the training of deep neural networks implicitly result in the minimization of $I(X;Z)$. In contrast, we studied the benefit of (implicitly or explicitly) controlling $I(X;Z)$.

In this paper, we consider *the generalization gap in deep learning* (Nagarajan & Kolter, 2019; Zhang et al., 2021a;b; Kawaguchi et al., 2022a;b; Hu et al., 2022), which is different from generalization gaps studied in the field of information bottleneck. In the field of information bottleneck, previous studies have analyzed a generalization gap between the true mutual information and its empirical estimate (Shamir et al., 2010; Tishby & Zaslavsky, 2015) and the generalization gap on $\mathcal{C}(q_\theta(Z|X), q(Y|Z))$ (Vera et al., 2018) where $\mathcal{C}(p,q)$ is the cross entropy of q relative to p, $q_\theta(Z|X)$ is a randomized encoder with learnable parameters θ, and $q(Y|Z)$ is a simple count-based decoder *with no learnable parameter*. Unlike ours, their bounds on $\mathcal{C}(q_\theta(Z|X), q(Y|Z))$ scale with $\frac{|\mathcal{X}| \ln n}{\sqrt{n}} + \frac{|\mathcal{Z}|}{\sqrt{n}}$ (due to their dependence on $\frac{1}{P_X(x_{\min})} \geq |\mathcal{X}|$ and $|\mathcal{U}| = |\mathcal{Z}|$ in their notation). This dependence on $|\mathcal{X}|$ makes their bounds inapplicable as it requires the number of samples $n \gg |\mathcal{X}|^2$. Here, the cross entropy $\mathcal{C}(q_\theta(Z|X), q(Y|Z))$ studied in the previous work is also different from the cross-entropy loss of deep learning, $\mathcal{C}(p(Y|X), q_\theta(Y|X))$, where $p(Y|X)$ is a target distribution and $q_\theta(Y|X)$ represents a deep neural network with learnable parameters θ. Therefore, we could not rely on any of these previous results from the field of information bottleneck.

Another related yet different topic of information theory is to use $I(f^S;S)$ to compute generalization bounds (Xu & Raginsky, 2017; Bassily et al., 2018; Hellström & Durisi, 2020; Steinke & Zakynthinou, 2020). However, these previous bounds are ***not*** about information bottleneck as these do ***not*** utilize $I(X;Z_l^s|Y)$ (or $I(X;Z_l^s)$) and only uses $I(f^S;S)$, the mutual information between the training dataset S and the ***entire*** model $f^S = \phi_{D+1}^S$. Thus, the previous bounds cannot provide insights or justifications on the information bottleneck principle unlike our bounds. Moreover, in Section 5, we demonstrate the advantage of $I(X;Z_l^s|Y) + I(\phi_l^S;S)$ in our bound over $I(f^S;S)$ in the previous bounds. Here, notice that $I(\phi_l^S;S) \neq I(f^S;S)$ for any $l \neq D+1$, and we always have $I(\phi_1^S;S) \leq \cdots \leq I(\phi_D^S;S) \leq I(\phi_{D+1}^S;S) = I(f^S;S)$ (e.g., see Fig. 2). See Appendix E.3 for more comparison with these previous bounds, which are ***not*** of information bottlenecks although they use mutual information.

The various interesting properties of information bottleneck are discussed in (Achille & Soatto, 2018). However, they do not provide generalization bounds with information bottlenecks. Among others, they discuss a connection between $\mathbf{I}(X;Z_l^s)$ and $\mathbf{I}(\tilde{\theta};S)$, where $\mathbf{I}$ represents an over-simplified version of mutual information in which everything is ignored except an artificially added noise; i.e., $\mathbf{I}(\tilde{\theta};S) \triangleq -\frac{1}{2}\sum_{i=1}^d \log \alpha_i$ (this definition is given in Remark 4.2 of Achille & Soatto, 2018) where α_i is the variance of the Gaussian noise ϵ that is artificially multiplied to the learned weights θ after training as $\tilde{\theta} = \epsilon\theta$. In other words, $\mathbf{I}$ completely ignores the dependence of θ on S, although it is the main factor in the question of generalization. For example, if we set $\alpha_i = 1$, we have $\mathbf{I}(\tilde{\theta};S) = 0$ always, despite the fact that $I(\theta;S) \neq 0$ when θ is learned from S. This shows that the previous paper did not consider the mutual information between θ and S; it used the entropy of the artificially multiplied Gaussian noise ϵ as mutual information. Thus the meaningful factors of mutual information are ignored in the previous work.

There are generalization bounds via different approaches including VC-dimension (Vapnik, 1999; Bartlett et al., 2019), Rademacher complexity (Bartlett & Mendelson, 2002; Truong, 2022; Träuble et al., 2023), stability (Bousquet & Elisseeff, 2002; Deng et al., 2021), robustness (Xu & Mannor, 2012; Liu et al., 2021; Kawaguchi et al., 2022a; Liu et al., 2022), and PAC-Bayes (Dziugaite et al., 2021; Lotfi et al., 2022). Unlike these previous studies, we provided the first generalization bound via the information bottleneck.

8. Conclusion

This study completed the proof of the previous conjecture with near-exponential improvements for the setting of fixed representations, proved the first rigorous generalization bound for the setting of learning representations, and empirically strengthened the findings with supporting experiments. This paper makes a contribution on technical aspects relevant for current and future methods of deep learning through the lens of information bottlenecks. Whereas information bottleneck is explicit in various algorithms (e.g., Federici et al., 2020; Sun et al., 2022; Li et al., 2022a;b; Su et al., 2023), it is also interesting to motivate future methods based on implicit effects of architectures (e.g., transformer and convolution) and training (e.g., SGD and self-supervised objectives) on information bottlenecks.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Amjad, R. A. and Geiger, B. C. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2225–2239, 2019.

Aziznejad, S., Gupta, H., Campos, J., and Unser, M. Deep neural networks with trainable activations and controlled lipschitz constant. *IEEE Transactions on Signal Processing*, 68:4688–4699, 2020.

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

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Bassily, R., Moran, S., Nachum, I., Shafer, J., and Yehudayoff, A. Learners that use little information. In *Algorithmic Learning Theory*, pp. 25–55. PMLR, 2018.

Bertsekas, D. P. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Burhanpurkar, M., Deng, Z., Dwork, C., and Zhang, L. Scaffolding sets. *arXiv preprint arXiv:2111.03135*, 2021.

Chelombiev, I., Houghton, C., and O'Donnell, C. Adaptive estimators show information compression in deep neural networks. In *International Conference on Learning Representations*, 2019.

Deng, Z., He, H., and Su, W. Toward better generalization bounds with locally elastic stability. In *International Conference on Machine Learning*, pp. 2590–2600. PMLR, 2021.

Dziugaite, G. K., Hsu, K., Gharbieh, W., Arpino, G., and Roy, D. On the role of data in pac-bayes bounds. In *International Conference on Artificial Intelligence and Statistics*, pp. 604–612. PMLR, 2021.

Eysenbach, B., Salakhutdinov, R. R., and Levine, S. Robust predictable control. *Advances in Neural Information Processing Systems*, 34:27813–27825, 2021.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020.

Fischer, I. The conditional entropy bottleneck. *Entropy*, 22 (9):999, 2020.

Galloway, A., Golubeva, A., Salem, M., Nica, M., Ioannou, Y., and Taylor, G. W. Bounding generalization error with input compression: An empirical study with infinite-width networks. *arXiv preprint arXiv:2207.09408*, 2022.

Goldfeld, Z., Van Den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., and Polyanskiy, Y. Estimating information flow in deep neural networks. In *International Conference on Machine Learning*, pp. 2299–2308. PMLR, 2019.

Goyal, A. and Bengio, Y. Inductive biases for deep learning of higher-level cognition. *Proc. A, Royal Soc., arXiv:2011.15091*, 2022.

Hafez-Kolahi, H., Kasaei, S., and Soleymani-Baghshah, M. Sample complexity of classification with compressed input. *Neurocomputing*, 415:286–294, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645. Springer, 2016.

Hellström, F. and Durisi, G. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1 (3):824–839, 2020.

Hromkovič, J. Randomized algorithms. In *Algorithmics for Hard Problems*, pp. 341–429. Springer, 2004.

Hu, Z., Jagtap, A. D., Karniadakis, G. E., and Kawaguchi, K. When do extended physics-informed neural networks (xpinns) improve generalization? *SIAM Journal on Scientific Computing*, 44(5):A3158–A3182, 2022.

Ji, W., Deng, Z., Nakada, R., Zou, J., and Zhang, L. The power of contrast for feature learning: A theoretical analysis. *arXiv*, 2110.02473, 2021a.

Ji, W., Lu, Y., Zhang, Y., Deng, Z., and Su, W. J. An unconstrained layer-peeled perspective on neural collapse. *arXiv*, 2110.02796, 2021b.

Kawaguchi, K., Deng, Z., Luh, K., and Huang, J. Robustness implies generalization via data-dependent generalization bounds. In *International Conference on Machine Learning (ICML)*, pp. 10866–10894. PMLR, 2022a.

Kawaguchi, K., Kaelbling, L., and Bengio, Y. *Generalization in Deep Learning*. Cambridge University Press, 2022b. URL https://arxiv.org/abs/1710.05468.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Kirsch, A., Lyle, C., and Gal, Y. Unpacking information bottlenecks: Surrogate objectives for deep learning. 2020.

Kolchinsky, A. and Tracey, B. D. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017.

Latorre, F., Rolland, P., and Cevher, V. Lipschitz constant estimation of neural networks via sparse polynomial optimization. In *International Conference on Learning Representations*, 2019.

Lee, K.-H., Arnab, A., Guadarrama, S., Canny, J., and Fischer, I. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34:19538–19552, 2021.

Levine, A. and Feizi, S. Robustness certificates for sparse adversarial attacks by randomized ablation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4585–4593, 2020.

Li, B., Shen, Y., Wang, Y., Zhu, W., Li, D., Keutzer, K., and Zhao, H. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7399–7407, 2022a.

Li, Q., Wu, Z., Kong, L., and Bi, W. Explanation regeneration via information bottleneck. *arXiv preprint arXiv:2212.09603*, 2022b.

Liu, D., Lamb, A., Kawaguchi, K., Goyal, A., Sun, C., Mozer, M. C., and Bengio, Y. Discrete-valued neural communication. In *Advances in Neural Information Processing Systems*, 2021.

Liu, D., Lamb, A., Ji, X., Notsawo, P., Mozer, M., Bengio, Y., and Kawaguchi, K. Adaptive discrete communication bottlenecks with dynamic vector quantization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Lotfi, S., Finzi, M., Kapoor, S., Potapczynski, A., Goldblum, M., and Wilson, A. G. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Pauli, P., Koch, A., Berberich, J., Kohler, P., and Allgöwer, F. Training robust neural networks using lipschitz bounds. *IEEE Control Systems Letters*, 6:121–126, 2021.

Pinot, R., Meunier, L., Araujo, A., Kashima, H., Yger, F., Gouy-Pailler, C., and Atif, J. Theoretical evidence for adversarial robustness through randomization. *Advances in Neural Information Processing Systems*, 32, 2019.

Pinot, R., Ettedgui, R., Rizk, G., Chevaleyre, Y., and Atif, J. Randomization matters how to defend against strong adversarial attacks. In *International Conference on Machine Learning*, pp. 7717–7727. PMLR, 2020.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shamir, O., Sabato, S., and Tishby, N. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Shwartz-Ziv, R., Painsky, A., and Tishby, N. Representation compression and generalization in deep neural networks, 2019. URL https://openreview.net/forum?id=SkeL6sCqK7.

Slonim, N. and Tishby, N. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 208–215, 2000.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

Su, T., Song, C., and Cheng, J. Vision transformer with information bottleneck for fine-grained visual classification. In *Advances in Guidance, Navigation and Control: Proceedings of 2022 International Conference on Guidance, Navigation and Control*, pp. 4010–4019. Springer, 2023.

Sun, Q., Li, J., Peng, H., Wu, J., Fu, X., Ji, C., and Philip, S. Y. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4165–4174, 2022.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, 2015.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pp. 368–377, 1999.

Träuble, F., Goyal, A., Rahaman, N., Mozer, M., Kawaguchi, K., Bengio, Y., and Schölkopf, B. Discrete key-value bottleneck. In *International Conference on Machine Learning (ICML)*, 2023.

Truong, L. V. On rademacher complexity-based generalization bounds for deep learning. *arXiv preprint arXiv:2208.04284*, 2022.

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999.

Vera, M., Piantanida, P., and Vega, L. R. The role of the information bottleneck in representation learning. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1580–1584. IEEE, 2018.

Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

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

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

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? In *International Conference on Learning Representations (ICLR)*, 2021b.

REPRODUCIBILITY STATEMENT

For the theoretical results, complete proofs are provided. For the empirical experiments, source code is available at: https://github.com/xu-ji/information-bottleneck

A. Experimental details for section 5.1

A.1. Training

Data. The dataset was 5-way classification on 2D clustered inputs (fig. 3). Each dataset draw contained 50 training points and 250 test points.

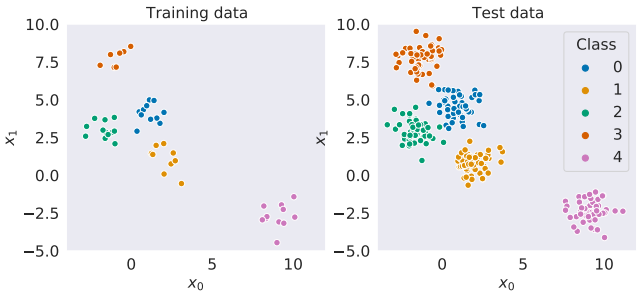


Figure 3. Example draw for 2D classification dataset.

Training. 216 models were trained for all combinations of options: 4 ReLU-activated MLP architectures (per-layer widths of [256, 256, 128, 128], [128, 128, 64, 64], [64, 64, 32, 32], [32, 32, 16, 16]), 3 weight decay rates (0, 0.01, 0.1), 3 dataset draws, 3 random seeds, and 2 sample set sizes for evaluating $I(X; Z_l^s)$ and $I(X; Z_l^s|Y)$. Final features were sampled from deterministically computed mean and standard deviation vectors and mapped to class probabilities with a softmax-activated linear layer. The expectation in MI over $\mathbb{P}(Z_l^s|x)$ depends on neural network parameters, so the reparameterization trick was used to optimize the expectation with respect to neural network parameters by rewriting the expectation over Z_l^s as an expectation over random noise (Kingma et al., 2015). Models were trained for 300 iterations with a learning rate of $\eta_\theta = 1e-2$. Out of all settings, 36 models with training set accuracy $< 85\%$ were discarded. Statistics for accepted models are given in table 4. Of 180 accepted models, 9 (5%) had a small negative generalization gap in loss (-0.0258 ± 0.0161). These models were not screened out before evaluating the metrics because the generalization gap is not estimatable without access to labelled test data.

	Mean	Standard deviation	Max	Min
Train loss	0.1265	0.1603	0.5757	0.0018
Train accuracy	0.9680	0.0479	1.0000	0.8600
Test loss	0.1984	0.1593	0.5487	0.0247
Test accuracy	0.9356	0.0568	0.9960	0.7880

Table 4. Performance statistics of 180 accepted models.

Constrained optimization. In each learning iteration, the gradient of the relaxed problem $\theta^s \leftarrow \theta^s - \eta_{\theta^s} \nabla_{\theta^s} \big[\big(- \frac{1}{|s|} \sum_{(x,y)\in s} \big(\log \frac{1}{k} \sum_{j=1}^{k} q_{\theta^s}(y|z^j) \big) \big) + \lambda(\rho - \hat{I}(X; Z_l^s)) \big]$ was applied to update the model and $\lambda \leftarrow \lambda + \eta_\lambda(\rho - \hat{I}(X; Z_l^s))$ was applied to update the multiplier λ, where η_{θ^s} and η_λ are learning rates. For a similar use case for dual gradient descent, see Eysenbach et al. (2021). Example plots showing the change in $\hat{I}(X; Z_l^s)$ and λ during training are given in fig. 4. Note that the gradient of λ is a term in the gradient of θ^s, thus updating λ incurs negligible additional cost.

A.2. Metrics

SWAG provides an estimate of the posterior as a multivariate Gaussian by averaging gradient updates across training epochs. SWAG was used in the estimator $\breve{I}(S; \theta_l^S) = (1/|D|) \sum_{s\in D} (1/k) \sum_{j=1}^{k} (\log p(w^j|s)) - ((1/|D|) \sum_{s'\in D} \log p(w^j|s')) \geq$

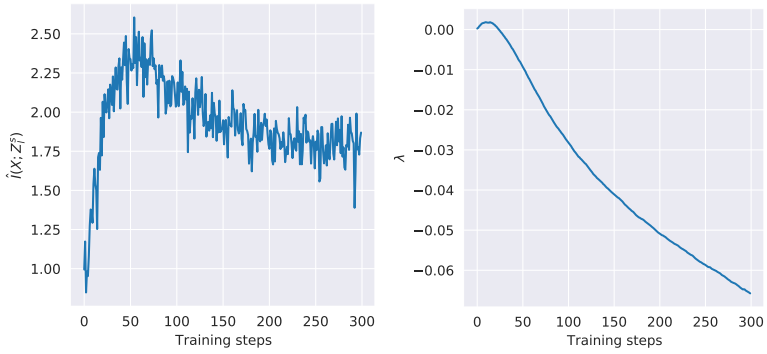


Figure 4. Example plots of $\hat{I}(X;Z_l^s)$ and λ during constrained optimized of a neural network model with $\rho = 1.5$.

$(1/|D|)\sum_{s\in D}(1/k)\sum_{j=1}^{k}\log(p(w^j|s)/((1/|D|)\sum_{s'\in D}p(w^j|s')))$, where $w^j \sim \mathbb{P}(\theta_l^S|S=s)$ for all j and the upper bound is obtained via Jensen's inequality. We found that averaging in the log domain by using the upper bound improved numerical stability compared to averaging in the probability domain due to large magnitudes of $\log p(w^j|s)$.

Mutual information between variables is a measure of their statistical dependence and is defined in our setting as:

$$I(X;Z_l^s) = \mathbb{E}_{X,Z_l^s}\log\frac{q_{\theta^s}(Z_l^s|X)}{\mathbb{E}_{X'}q_{\theta^s}(Z_l^s|X')}, \tag{8}$$

$$I(X;Z_l^s|Y) = \mathbb{E}_Y\mathbb{E}_{X_Y,Z_l^s}\log\frac{q_{\theta^s}(Z_l^s|X_Y)}{\mathbb{E}_{X'_Y}q_{\theta^s}(Z_l^s|X'_Y)}, \tag{9}$$

$$I(S;\theta_l^S) = \mathbb{E}_S\mathbb{E}_{\theta_l^S|S}\log\frac{\mathbb{P}(\theta_l^S|S)}{\mathbb{E}_{S''}\mathbb{P}(\theta_l^{S''}|S'')}, \tag{10}$$

$$\tag{11}$$

where X', X'_Y, S'' are independent copies of variables X, X_Y, S respectively. Let C be the set of classes and s_c denote dataset samples for class c. We use $\hat{I}$ to denote estimation by Monte-Carlo sampling and $\breve{I}$ to denote upper bounding via the Jensen inequality:

$$\hat{I}(X;Z_l^s) = \frac{1}{|s|}\sum_{(x,y)\in s}\frac{1}{k}\sum_{j=1}^{k}\log\frac{q_{\theta^s}(z^j|x)}{\frac{1}{|s|}\sum_{(x',y')\in s}q_{\theta^s}(z^j|x')}, \tag{12}$$

$$\hat{I}(X;Z_l^s|Y) = \frac{1}{|s|}\sum_{c\in C}\sum_{(x,y)\in s_c}\frac{1}{k}\sum_{j=1}^{k}\log\frac{q_{\theta^s}(z^j|x)}{\frac{1}{|s_c|}\sum_{(x',y')\in s_c}q_{\theta^s}(z^j|x')}, \tag{13}$$

$$\breve{I}(X;Z_l^s) = \frac{1}{|s|}\sum_{(x,y)\in s}\frac{1}{k}\sum_{j=1}^{k}\left(\log q_{\theta^s}(z^j|x) - \left(\frac{1}{|s|}\sum_{(x',y')\in s}\log q_{\theta^s}(z^j|x')\right)\right), \tag{14}$$

$$\breve{I}(X;Z_l^s|Y) = \frac{1}{|s|}\sum_{c\in C}\sum_{(x,y)\in s_c}\frac{1}{k}\sum_{j=1}^{k}\left(\log q_{\theta^s}(z^j|x) - \left(\frac{1}{|s_c|}\sum_{(x',y')\in s_c}\log q_{\theta^s}(z^j|x')\right)\right), \tag{15}$$

where $z^j \sim q_{\theta^s}(Z_l^s|x)$ for all j.

For mutual information between the model and training dataset, we compute:

$$\breve{I}(S;\theta_l^S) = \frac{1}{|D|}\sum_{s\in D}\frac{1}{k}\sum_{j=1}^{k}\left(\log p(w^j|s) - \left(\frac{1}{|D|}\sum_{s'\in D}\log p(w^j|s')\right)\right), \tag{16}$$

where $w^j \sim \mathbb{P}(\theta_l^S|S=s)$. The learning algorithm is defined by the variables excluding the training dataset, i.e. architecture, weight decay, multiplier learning rate, seed. Denote the average values of $\breve{I}(S;\theta_l^S)$ and $\hat{I}(X;Z_l^s|Y)$ across learning algorithms by μ and μ' respectively. The rescaled value $\tilde{I}(S;\theta_l^S)$ is defined by:

$$\tilde{I}(S;\theta_l^S) = \frac{\mu'}{\mu}\breve{I}(S;\theta_l^S). \tag{17}$$

Note that MI is measured in universal units, but we test multiple estimation procedures for $I(X;Z_l^s)$ and $I(S;\theta_l^S)$. Scaling was tested because of the difference of estimators, rather than of units of the estimated quantity.

A.3. Additional results

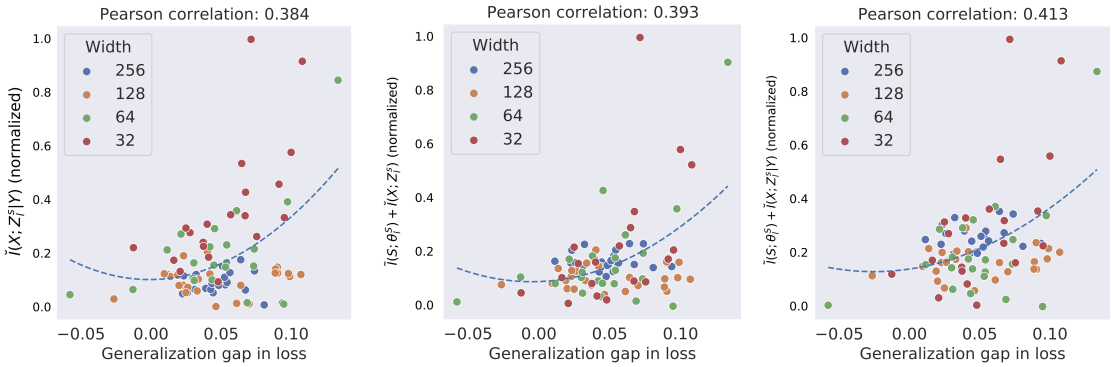
As the true data generator is available for the toy dataset, 2 sample set sizes were considered for computing estimators of $I(X;Z_l^s)$ and $I(X;Z_l^s|Y)$ during evaluation of the metrics (appendix A.1): using the training dataset (50 data points), and using a sample 10 times larger drawn from the generator (500 data points). Using the larger sample size (tables 1 and 5) improved the predictive ability of the baseline representation compression metrics compared to using the small sample size (table 6).

Metric	Spearman	Pearson	Kendall
Num. params. m	-0.0576	-0.0294	-0.0402
$m \log m$	-0.0576	-0.0287	-0.0402
$\sum_{\mathbf{l}} \theta_{\mathbf{l}}^s$	-0.2550	-0.1366	-0.1567
$\prod_{\mathbf{l}} \theta_{\mathbf{l}}^s$	-0.2172	-0.0871	-0.1374
$\hat{I}(X;Z_l^s)$	0.1816	0.2878	0.1280
$\hat{I}(X;Z_l^s\|Y)$	0.1749	0.3167	0.1129
$\breve{I}(X;Z_l^s)$	0.1648	0.3712	0.1223
$\breve{I}(X;Z_l^s\|Y)$	0.2293	0.3842	0.1515
$\breve{I}(S;\theta_l^S)$	0.0020	0.0211	0.0074
$\breve{I}(S;\theta_{D+1}^S)$	-0.0221	0.0091	-0.0090
$\breve{I}(S;\theta_l^S) + \hat{I}(X;Z_l^s)$	0.0178	0.0211	0.0178
$\breve{I}(S;\theta_l^S) + \hat{I}(X;Z_l^s\|Y)$	0.0163	0.0211	0.0167
$\breve{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s)$	0.0135	0.0212	0.0162
$\breve{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s\|Y)$	0.0164	0.0211	0.0167
$\tilde{I}(S;\theta_l^S) + \hat{I}(X;Z_l^s)$	0.1104	0.1401	0.0794
$\tilde{I}(S;\theta_l^S) + \hat{I}(X;Z_l^s\|Y)$	0.2253	0.3177	0.1567
$\tilde{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s)$	0.2684	0.3928	0.1912
$\tilde{I}(S;\theta_l^S) + \breve{I}(X;Z_l^s\|Y)$	**0.3015**	**0.4130**	**0.2085**

Table 5. Correlation coefficients for metrics and the generalization gap in loss, large sample setting for estimation of $I(X;Z_l^s)$ and $I(X;Z_l^s|Y)$. $\theta_{\mathbf{l}}^S$ denotes parameters of layer $\mathbf{l}$ and θ_l^S denotes parameters up to layer l. Layer l is fixed to the penultimate layer. > 0 indicates positive correlation.

Metric	Spearman	Pearson	Kendall
$\hat{I}(X;Z_l^s)$	-0.1066	-0.0972	-0.0709
$\hat{I}(X;Z_l^s\|Y)$	0.0868	0.0394	0.0698
$\breve{I}(S;\theta_l^S) + \hat{I}(X;Z_l^s\|Y)$	0.2360	0.2489	0.1611
$\tilde{I}(S;\theta_l^S) + \hat{I}(X;Z_l^s\|Y)$	**0.3277**	**0.2888**	**0.2257**

Table 6. Correlation coefficients for metrics and the generalization gap in loss, small sample setting for estimation of $I(X;Z_l^s)$ and $I(X;Z_l^s|Y)$. Best metric and ablation shown. Layer l is fixed to the penultimate layer. > 0 indicates positive correlation.



(a) $\breve{I}(X;Z_l^s|Y)$ (b) $\tilde{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$ (c) $\tilde{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s|Y)$

Figure 5. Correlation of best 3 metrics with the generalization gap. Color denotes network width. Dashed line denotes best polynomial fit with degree 2. Values are normalized by subtracting the minimum and dividing by the range.

B. MNIST and Fashion MNIST

We conducted experiments on the MNIST and Fashion MNIST datasets. These experiments follow the same protocol described in section 5.1 and appendix A except that MI was not constrained, in order to investigate predictive ability of the metrics in the setting of unconstrained stochastic feature models. 144 models were trained over all combinations of options: 2 datasets, 3 ReLU-activated architectures (1 convolutional layer and 3 linear layers, with per hidden layer channel sizes of [64, 512, 512], [32, 256, 256], [16, 128, 128]), 2 weight decay rates (0, 1e-3), 2 batch sizes (128, 32), 3 dataset draws, 2 random seeds. As in section 5.1 and appendix A, the penultimate layer of the network infers mean and standard deviation vectors that define a distribution over latent features. To construct multiple instances of the training dataset, we sampled training datasets of size 8K from the training set, and each test set was the original 10K test set. Performance of trained models are given in tables 7 and 8. In line with sections 5.1 and 5.2 and appendices A and D, results on MNIST and Fashion MNIST indicate that metrics which combine model compression with representation compression outperform metrics for representation compression alone (tables 9 and 10).

	Mean	Standard deviation	Max	Min
Train loss	0.0071	0.0165	0.1053	0.0001
Train accuracy	0.9986	0.0060	1.0000	0.9628
Test loss	0.1355	0.0285	0.2564	0.0915
Test accuracy	0.9673	0.0065	0.9737	0.9358

Table 7. Performance statistics for MNIST models.

	Mean	Standard deviation	Max	Min
Train loss	0.0692	0.0640	0.1908	0.0003
Train accuracy	0.9765	0.0234	1.0000	0.9329
Test loss	0.5609	0.1682	0.8819	0.3791
Test accuracy	0.8721	0.0090	0.8864	0.8262

Table 8. Performance statistics for Fashion MNIST models.

Metric	Spearman	Pearson	Kendall
$\hat{I}(X;Z_l^s)$	0.2738	-0.1268	0.2178
$\hat{I}(X;Z_l^s\|Y)$	0.4399	0.2059	0.3243
$\breve{I}(X;Z_l^s)$	0.7895	0.5467	0.6346
$\breve{I}(X;Z_l^s\|Y)$	0.7931	0.5348	**0.6416**
$\breve{I}(S;\theta_l^S)$	0.6004	0.7044	0.4193
$\bar{I}(S;\theta_l^S)$	0.5328	0.6906	0.3771
$\breve{I}(S;\theta_{D+1}^S)$	0.5384	0.6619	0.3768
$\bar{I}(S;\theta_{D+1}^S)$	0.5328	0.6507	0.3771
$\tilde{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	0.6021	0.7043	0.4130
$\tilde{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	0.5958	0.7044	0.3955
$\tilde{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	**0.8352**	0.6367	**0.6452**
$\tilde{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	**0.8303**	0.6242	0.6384
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	0.6021	0.7044	0.4130
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	0.5958	0.7044	0.3955
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	0.7128	**0.7626**	0.5119
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	0.6566	**0.7329**	0.4610

Table 9. MNIST. Correlation coefficients for metrics and the generalization gap in loss. Layer l is fixed to the penultimate layer. > 0 indicates positive correlation.

Metric	Spearman	Pearson	Kendall
$\hat{I}(X;Z_l^s)$	-0.0299	-0.2056	-0.0261
$\hat{I}(X;Z_l^s\|Y)$	0.2318	0.1185	0.1458
$\breve{I}(X;Z_l^s)$	0.3861	0.5146	0.2293
$\breve{I}(X;Z_l^s\|Y)$	0.3848	0.5115	0.2308
$\breve{I}(S;\theta_l^S)$	0.3479	0.3191	0.2682
$\bar{I}(S;\theta_l^S)$	0.3471	0.2804	0.2684
$\breve{I}(S;\theta_{D+1}^S)$	0.3479	0.3187	0.2682
$\bar{I}(S;\theta_{D+1}^S)$	0.3928	0.3121	0.2998
$\tilde{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	0.3377	0.3188	0.2469
$\tilde{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	0.3446	0.3191	0.2660
$\tilde{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	**0.5623**	**0.6488**	**0.4238**
$\tilde{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	**0.5637**	**0.6441**	**0.4344**
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	0.3377	0.3191	0.2469
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	0.3446	0.3191	0.2660
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	0.3734	0.3280	0.2677
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	0.3679	0.3217	0.2766

Table 10. Fashion MNIST. Correlation coefficients for metrics and the generalization gap in loss. Layer l is fixed to the penultimate layer. > 0 indicates positive correlation.

C. Feature binning

Binning is a method for characterizing feature compression in neural networks with deterministic features (Shwartz-Ziv & Tishby, 2017). To investigate the performance of feature and model compression for predicting generalization, we trained models for the toy clustering problem of section 5.1 using MLPs with deterministic features. The binning implementation of Saxe et al. (2019) was reused to discretize the activity of each node into 10 buckets. 216 models were trained over all

combinations of: 3 ReLU-activated MLP acrhitectures (5 fully connected layers with per hidden layer channel sizes of [512, 512, 256, 256], [256, 256, 128, 128], [128, 128, 64, 64]), 3 weight decay rates (0, 1e-3, 1e-2), 3 dataset draws, 3 random seeds, with and without information bottleneck (IB) feature regularization. For models trained with IB regularization, we used the surrogate objective of Kirsch et al. (2020, Eq. 1). While feature compression and model compression both performed reasonably in the setting without IB regularization (table 11), utilizing IB regularization significantly impaired the performance of metrics based solely on feature compression (table 12).

Metric	Spearman	Pearson	Kendall
$\hat{I}(X; Z_l^s)$	**0.3271**	**0.3032**	**0.2235**
$\hat{I}(X; Z_l^s \mid Y)$	**0.2431**	0.2194	**0.1643**
Num. param. m	0.0078	-0.0259	0.0064
$m \log m$	0.0078	-0.0267	0.0064
$\sum_l \theta_l^s$	0.0557	0.0770	0.0444
$\prod_l \theta_l^s$	-0.0559	-0.1755	-0.0191
$\breve{I}(S; \theta_l^S)$	0.1955	0.1851	0.1335
$\bar{I}(S; \theta_l^S)$	0.1955	0.1940	0.1347
$\tilde{I}(S; \theta_l^S) + \hat{I}(X; Z_l^s)$	0.2046	**0.2419**	0.1423
$\tilde{I}(S; \theta_l^S) + \hat{I}(X; Z_l^s \mid Y)$	0.1846	0.2389	0.1272
$\breve{I}(S; \theta_l^S) + \hat{I}(X; Z_l^s)$	0.1530	0.2041	0.1043
$\breve{I}(S; \theta_l^S) + \hat{I}(X; Z_l^s \mid Y)$	0.1377	0.2041	0.0920

Table 11. Results for clustering experiments using feature binning for estimation of MI, without information bottleneck regularization. Correlation coefficients for metrics and the generalization gap in loss. Layer l is fixed to the penultimate layer. > 0 indicates positive correlation.

Metric	Spearman	Pearson	Kendall
$\hat{I}(X; Z_l^s)$	-0.0403	-0.0440	-0.0426
$\hat{I}(X; Z_l^s \mid Y)$	0.0252	0.0390	-0.0012
Num. param. m	-0.0103	0.0057	-0.0229
$m \log m$	-0.0103	0.0053	-0.0229
$\sum_l \theta_l^s$	-0.0579	0.0158	-0.0457
$\prod_l \theta_l^s$	-0.1460	-0.1640	-0.1111
$\breve{I}(S; \theta_l^S)$	0.0252	0.0567	0.0159
$\bar{I}(S; \theta_l^S)$	0.0370	0.0665	0.0208
$\tilde{I}(S; \theta_l^S) + \hat{I}(X; Z_l^s)$	0.0498	0.0545	0.0381
$\tilde{I}(S; \theta_l^S) + \hat{I}(X; Z_l^s \mid Y)$	0.0688	0.0647	0.0512
$\breve{I}(S; \theta_l^S) + \hat{I}(X; Z_l^s)$	**0.0757**	**0.0815**	**0.0530**
$\breve{I}(S; \theta_l^S) + \hat{I}(X; Z_l^s \mid Y)$	**0.0794**	**0.0815**	**0.0525**

Table 12. Results for clustering experiments using feature binning for estimation of MI, with information bottleneck regularization. Correlation coefficients for metrics and the generalization gap in loss. Layer l is fixed to the penultimate layer. > 0 indicates positive correlation.

D. Experimental details for section 5.2

D.1. Training

540 models were trained for all combinations of the options: 3 architecutres (PreResNet56, PreResNet83, PreResNet110), 3 weight decay rates (1e-3, 1e-4, 1e-5), 3 batch sizes (64, 128, 1024), 5 dataset draws and 4 random seeds. The PreResNet architecture (He et al., 2016) consists of a convolutional layer, 3 residual blocks and a final linear prediction layer. We consider representations from $D = 5$ layers: input layer, after convolutional layer, and after each residual block. Models

were trained for 200 epochs with SGD and a learning rate of $1e-2$. Statistics are given in table 13.

	Mean	Standard deviation	Max	Min
Train loss	0.2157	0.2815	1.4174	0.0005
Train accuracy	0.9282	0.0937	1.0000	0.5433
Test loss	0.9704	0.2023	1.6269	0.5989
Test accuracy	0.8043	0.0554	0.8770	0.5192

Table 13. Performance statistics of 540 models.

D.2. Metrics

We used the same metrics as defined in appendix A.2 and additionally test excluding the seed from the definition of the learning algorithm by averaging across seeds. Let G denote the set of seeds and let Γ denote the seed variable:

$$\bar{I}(S;\theta_l^S) = \frac{1}{|D||G|}\sum_{s\in D}\sum_{\gamma\in G}\frac{1}{k}\sum_{j=1}^{k}\left(\left(\frac{1}{|G|}\sum_{\gamma'\in G}\log p(w^j|s,\gamma')\right) - \left(\frac{1}{|D||G|}\sum_{s'\in D}\sum_{\gamma'\in G}\log p(w^j|s',\gamma')\right)\right) \tag{18}$$

where $w^j \sim \mathbb{P}(\theta_l^S|S=s,\Gamma=\gamma)$ is sampled from the estimated posterior produced by SWAG.

D.3. Kernel Density Estimation

Without the addition of noise in the hidden representation, mutual information between inputs and deterministic continuous features is ill-defined (Saxe et al., 2019). One way to add noise is to discretize hidden activity into bins (Shwartz-Ziv & Tishby, 2017; Burhanpurkar et al., 2021). Another approach is kernal density estimation (Kolchinsky & Tracey, 2017; Ji et al., 2021a;b), which assumes for the purpose of analysis that Gaussian noise with variance σ_l^2 is added to the representation produced by layer l. In adaptive KDE (Chelombiev et al., 2019) σ_l^2 is scaled from a base by the maximum observed activation level in the layer, improving on constant σ_l^2 (Saxe et al., 2019) by allowing the level of noise to vary with layers. Following the previous work, we found $1e-3$ to work well as the base value.

As an alternative method for specifying σ_l^2, we selected σ_l^2 from a discrete set by maximum log likelihood of observed features,

$$\frac{1}{|s|}\sum_{(x,y)\in s}\frac{1}{k}\sum_{j=1}^{k}\log\frac{1}{|s|}\sum_{(x',y')\in s}q_{\theta^s}(z^j|x') \qquad \text{where } z^j \sim q_{\theta^s}(Z_l^s|x), \tag{19}$$

under the constraint that estimated MI decreased with layer, which follows from the information processing inequality. This was performed by iterating from layer D to layer 1 and choosing σ_l^2 with maximum likelihood such that the estimator of MI was non-decreasing, i.e. $\hat{I}(X;Z_l^s) \geq \hat{I}(X;Z_{l+1}^s)$ for $l < D$. As with estimators in appendix A.2, averaging can be done in the log domain to yield the lower bound:

$$\frac{1}{|s|}\sum_{(x,y)\in s}\frac{1}{k}\sum_{j=1}^{k}\frac{1}{|s|}\sum_{(x',y')\in s}\log q_{\theta^s}(z^j|x') \qquad \text{where } z^j \sim q_{\theta^s}(Z_l^s|x). \tag{20}$$

For consistency, eq. (19) was used with $\hat{I}(X;Z_l^s)$ (eq. (12)) and eq. (20) with $\breve{I}(X;Z_l^s)$ (eq. (14)).

D.4. Further results

We provide further results below.

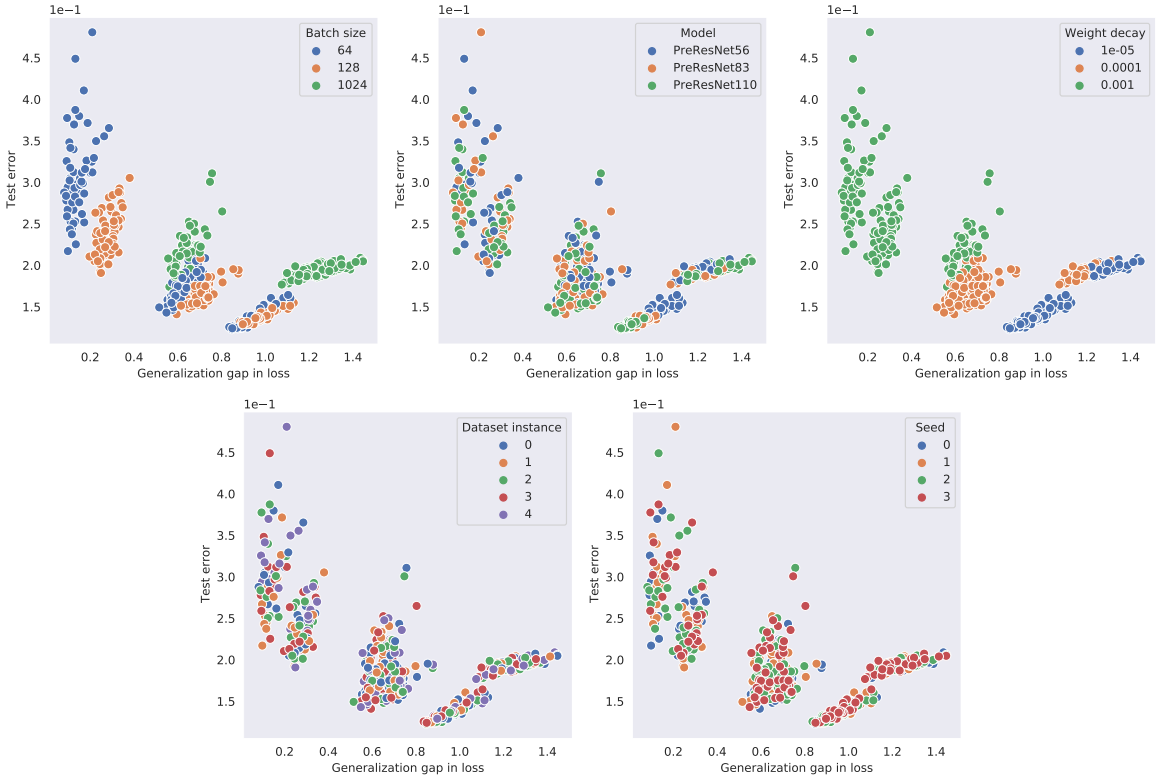


Figure 6. Illustration of performance-based clustering behaviour that emerged from training, attributed mostly to batch size and weight decay.

	Spearman corr.		Pearson corr.		Kendall corr.	
Generalization gap:	Loss	Error	Loss	Error	Loss	Error
Train loss	-0.8075	-0.7229	-0.8376	-0.8730	-0.5632	-0.5307
Test loss	0.6970	0.6095	0.6497	0.5210	0.5938	0.4777
Train error	-0.8114	-0.7269	-0.8489	-0.8831	-0.5707	-0.5381
Test error	-0.4609	-0.3834	-0.5896	-0.6087	-0.2652	-0.2020
Gen. gap error	0.9443	1.0000	0.9562	1.0000	0.8149	1.0000
Gen. gap loss	1.0000	0.9443	1.0000	0.9562	1.0000	0.8149

Table 14. Results for prediction with performance metrics.

Layer index l:	1	2	3	4	5
$\breve{I}(X;Z_l^s)$	5.9422E+04	1.8236E+04	1.5514E+04	7.7865E+03	5.7027E+03
$\breve{I}(X;Z_l^s\|Y)$	5.7783E+04	1.7790E+04	1.5429E+04	7.6948E+03	5.3298E+03
$\bar{I}(S;\theta_l^S)$	0.0000E+00	3.7136E+03	9.4890E+05	2.1247E+06	5.8244E+06
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	5.7783E+04	2.1504E+04	9.6433E+05	2.1324E+06	5.8297E+06

Table 15. Example values of metrics, for best performing model by test loss (PreResNet56, batch size 128, weight decay 0.001). $l=1$ is the input layer.

Generalization gap:	Spearman corr. Loss	Spearman corr. Error	Pearson corr. Loss	Pearson corr. Error	Kendall corr. Loss	Kendall corr. Error
$\breve{I}(S;\theta^S_{D+1})$	0.4688	0.3112	0.2512	0.0775	0.2121	0.1208
$\bar{I}(S;\theta^S_{D+1})$	0.5370	0.3800	0.2924	0.1218	0.2442	0.1526

Table 16. Results for model compression metrics.

Generalization gap:	Layer	Spearman corr. Loss	Spearman corr. Error	Pearson corr. Loss	Pearson corr. Error	Kendall corr. Loss	Kendall corr. Error
$\hat{I}(X;Z^s_l)$	$l=1$	0.7345	0.6156	0.5722	0.3853	0.5174	0.4215
$\hat{I}(X;Z^s_l)$	$l=D$	0.7170	0.5740	0.7063	0.5602	0.4784	0.3721
$\hat{I}(X;Z^s_l\|Y)$	$l=1$	0.7199	0.6073	0.5724	0.3854	0.5005	0.4111
$\hat{I}(X;Z^s_l\|Y)$	$l=D$	0.7126	0.5691	0.7071	0.5616	0.4768	0.3700
$\breve{I}(X;Z^s_l)$	$l=1$	0.6765	0.5655	0.1554	0.1328	0.4553	0.3781
$\breve{I}(X;Z^s_l)$	$l=D$	0.7145	0.5602	0.7203	0.5719	0.4461	0.3404
$\breve{I}(X;Z^s_l\|Y)$	$l=1$	0.6476	0.5292	0.1557	0.1331	0.4307	0.3504
$\breve{I}(X;Z^s_l\|Y)$	$l=D$	0.7004	0.5434	0.7062	0.5560	0.4386	0.3305

Table 17. Results for representation compression metrics for $l \in \{1, D\}$ summarization over layers (MLE selection of σ_l^2).

Generalization gap:	Layer	Spearman corr. Loss	Spearman corr. Error	Pearson corr. Loss	Pearson corr. Error	Kendall corr. Loss	Kendall corr. Error
$\hat{I}(X;Z^s_l)$	$l=1$	-0.0681	-0.0659	-0.0645	-0.0617	-0.0464	-0.0447
$\hat{I}(X;Z^s_l)$	$l=D$	0.7019	0.5452	0.4596	0.2845	0.4329	0.3293
$\hat{I}(X;Z^s_l\|Y)$	$l=1$	0.0050	0.0032	0.0146	0.0191	0.0035	0.0022
$\hat{I}(X;Z^s_l\|Y)$	$l=D$	0.6977	0.5403	0.4544	0.2798	0.4302	0.3268
$\breve{I}(X;Z^s_l)$	$l=1$	-0.0196	-0.0080	-0.0070	-0.0029	-0.0139	-0.0051
$\breve{I}(X;Z^s_l)$	$l=D$	0.7314	0.5848	0.5109	0.3350	0.4645	0.3624
$\breve{I}(X;Z^s_l\|Y)$	$l=1$	0.0005	-0.0060	0.0036	-0.0001	0.0005	-0.0038
$\breve{I}(X;Z^s_l\|Y)$	$l=D$	0.7033	0.5465	0.4587	0.2846	0.4342	0.3304

Table 18. Results for metrics for $l \in \{1, D\}$ summarization over layers (adaptive KDE selection of σ_l^2).

Generalization gap:	Layer	Spearman corr. Loss	Spearman corr. Error	Pearson corr. Loss	Pearson corr. Error	Kendall corr. Loss	Kendall corr. Error
$\hat{I}(X;Z_l^s)$	Mean	0.7598	0.6184	0.5781	0.3911	0.4970	0.3936
$\hat{I}(X;Z_l^s)$	Max	0.7345	0.6156	0.5722	0.3853	0.5174	0.4215
$\hat{I}(X;Z_l^s)$	Min	0.7170	0.5740	0.7063	0.5602	0.4784	0.3721
$\hat{I}(X;Z_l^s\|Y)$	Mean	0.7720	0.6378	0.5790	0.3920	0.5186	0.4145
$\hat{I}(X;Z_l^s\|Y)$	Max	0.7049	0.5814	0.5723	0.3853	0.4712	0.3785
$\hat{I}(X;Z_l^s\|Y)$	Min	0.7137	0.5701	0.7112	0.5679	0.4775	0.3697
$\breve{I}(X;Z_l^s)$	Mean	0.8481	0.7410	0.2116	0.1831	0.6425	0.5436
$\breve{I}(X;Z_l^s)$	Max	0.6765	0.5655	0.1554	0.1328	0.4553	0.3781
$\breve{I}(X;Z_l^s)$	Min	0.7145	0.5602	0.7203	0.5719	0.4461	0.3404
$\breve{I}(X;Z_l^s\|Y)$	Mean	0.8481	0.7406	0.2140	0.1853	0.6427	0.5435
$\breve{I}(X;Z_l^s\|Y)$	Max	0.6486	0.5297	0.1557	0.1331	0.4316	0.3511
$\breve{I}(X;Z_l^s\|Y)$	Min	0.7004	0.5434	0.7062	0.5560	0.4386	0.3305
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Mean	0.4546	0.3055	0.1867	0.0160	0.2304	0.1398
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Max	0.5111	0.3609	0.2572	0.0858	0.2634	0.1715
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Min	0.8134	0.6906	0.5363	0.3729	0.5840	0.4870
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Mean	0.4543	0.3052	0.1858	0.0152	0.2305	0.1397
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Max	0.5112	0.3609	0.2572	0.0858	0.2637	0.1718
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Min	0.8136	0.6949	0.5193	0.3591	0.5817	0.4871
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Mean	0.4513	0.3026	0.1832	0.0132	0.2305	0.1398
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Max	0.5112	0.3609	0.2572	0.0858	0.2636	0.1715
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Min	0.8489	0.7353	0.8459	0.7216	0.6354	0.5386
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Mean	0.4513	0.3026	0.1832	0.0132	0.2301	0.1397
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Max	0.5113	0.3609	0.2572	0.0858	0.2638	0.1716
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Min	0.8434	0.7313	0.8437	0.7195	0.6270	0.5332
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Mean	0.4770	0.3244	0.2901	0.1155	0.2510	0.1568
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Max	0.5709	0.4205	0.2993	0.1311	0.2878	0.1946
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Min	0.7898	0.6548	0.6706	0.4929	0.5432	0.4412
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Mean	0.4764	0.3241	0.2869	0.1128	0.2489	0.1558
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Max	0.5709	0.4205	0.2993	0.1311	0.2882	0.1952
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Min	0.7917	0.6595	0.6490	0.4753	0.5447	0.4450
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Mean	0.4429	0.2910	0.2785	0.1060	0.2353	0.1432
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Max	0.5707	0.4205	0.2993	0.1311	0.2880	0.1946
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Min	**0.8635**	**0.7576**	**0.8493**	**0.7544**	**0.6660**	**0.5684**
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Mean	0.4429	0.2908	0.2783	0.1059	0.2349	0.1426
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Max	0.5711	0.4204	0.2993	0.1311	0.2886	0.1945
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Min	**0.8632**	**0.7576**	**0.8511**	**0.7562**	**0.6626**	**0.5664**

Table 19. Results for metrics for mean, min, max summarization over layers (MLE selection of σ_l^2). Best metrics highlighted.

Generalization gap:	Layer	Spearman corr. Loss	Spearman corr. Error	Pearson corr. Loss	Pearson corr. Error	Kendall corr. Loss	Kendall corr. Error
$\hat{I}(X;Z_l^s)$	Mean	0.7322	0.5837	0.6823	0.5067	0.4606	0.3605
$\hat{I}(X;Z_l^s)$	Max	0.7805	0.6521	0.6848	0.5088	0.5299	0.4360
$\hat{I}(X;Z_l^s)$	Min	0.7726	0.6450	0.8131	0.7004	0.5126	0.4205
$\hat{I}(X;Z_l^s\|Y)$	Mean	0.7323	0.5833	0.6863	0.5110	0.4604	0.3598
$\hat{I}(X;Z_l^s\|Y)$	Max	0.7865	0.6548	0.6920	0.5160	0.5378	0.4380
$\hat{I}(X;Z_l^s\|Y)$	Min	0.7731	0.6451	0.8175	0.7041	0.5131	0.4221
$\breve{I}(X;Z_l^s)$	Mean	0.7401	0.5933	0.7078	0.5334	0.4739	0.3727
$\breve{I}(X;Z_l^s)$	Max	0.7452	0.6264	0.6891	0.5116	0.4927	0.4109
$\breve{I}(X;Z_l^s)$	Min	0.7981	0.6720	**0.8515**	**0.7262**	0.5490	0.4500
$\breve{I}(X;Z_l^s\|Y)$	Mean	0.7375	0.5894	0.7010	0.5263	0.4701	0.3685
$\breve{I}(X;Z_l^s\|Y)$	Max	0.7449	0.6177	0.7056	0.5289	0.4924	0.4053
$\breve{I}(X;Z_l^s\|Y)$	Min	0.7686	0.6270	0.8130	0.6699	0.5093	0.4055
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Mean	0.4555	0.3070	0.1856	0.0153	0.2334	0.1425
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Max	0.5133	0.3633	0.2580	0.0865	0.2650	0.1735
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Min	0.8112	0.7033	0.8272	0.7243	0.5775	0.4914
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Mean	0.4542	0.3058	0.1849	0.0146	0.2322	0.1414
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Max	0.5126	0.3625	0.2578	0.0863	0.2643	0.1730
$\breve{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Min	**0.8205**	**0.7203**	0.8281	**0.7313**	**0.5913**	**0.5108**
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Mean	0.4602	0.3115	0.1878	0.0172	0.2378	0.1468
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Max	0.5146	0.3647	0.2588	0.0871	0.2666	0.1748
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Min	0.8014	0.6854	0.8423	0.7107	0.5658	0.4760
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Mean	0.4598	0.3111	0.1874	0.0168	0.2374	0.1464
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Max	0.5143	0.3643	0.2584	0.0868	0.2661	0.1743
$\breve{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Min	0.8069	0.6913	**0.8440**	0.7135	0.5732	0.4849
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Mean	0.4783	0.3249	0.2868	0.1134	0.2495	0.1577
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Max	0.5765	0.4250	0.3021	0.1333	0.2971	0.2057
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s)$	Min	0.8031	0.6920	0.8270	0.7149	0.5664	0.4793
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Mean	0.4766	0.3242	0.2842	0.1110	0.2468	0.1551
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Max	0.5759	0.4247	0.3013	0.1326	0.2958	0.2049
$\bar{I}(S;\theta_l^S)+\hat{I}(X;Z_l^s\|Y)$	Min	**0.8130**	**0.7070**	0.8295	0.7203	**0.5799**	**0.4963**
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Mean	0.4864	0.3335	0.2946	0.1205	0.2613	0.1689
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Max	0.5769	0.4253	0.3048	0.1356	0.2978	0.2067
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s)$	Min	0.7974	0.6749	0.8206	0.6842	0.5608	0.4684
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Mean	0.4851	0.3317	0.2933	0.1192	0.2596	0.1674
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Max	0.5770	0.4253	0.3034	0.1343	0.2981	0.2068
$\bar{I}(S;\theta_l^S)+\breve{I}(X;Z_l^s\|Y)$	Min	0.7989	0.6763	0.8242	0.6891	0.5628	0.4703

Table 20. Results for metrics for mean, min, max summarization over layers (adaptive KDE selection of σ_l^2). Best metrics highlighted.

E. Additional Results and Explanations for Theory

We present additional results and explanations for theoretical results in appendix E, full proofs in appendix F, and additional results and details for experimental results in appendix A.

E.1. On Theorems 1–2

The mutual information $I(\phi_l^S; S)$ in Theorem 2 does not appear in Conjecture 1. However, the sample complexity bound in Conjecture 1 is invalid for the setting of learning ϕ_l^s, because the encoder ϕ_l^s can overfit to the training data, which was demonstrated with the counter-example in Hafez-Kolahi et al. (2020, Example 3.1). The mutual information $I(\phi_l^S; S)$ is measuring the effect of overfitting the encoder, which is necessary to avoid the counter-example.

Using additional notation defined in appendix E.1.1, we will discuss the details of the formulas of $G_1^l(0)$, $\mathcal{G}_2^l$ and G_3^l of Theorems 1 and $G_1^l(\zeta)$, $\widehat{\mathcal{G}}_2^l$, $\check{\mathcal{G}}_2^l$, and G_3^l of Theorem 2 in appendix E.1.2.

E.1.1. Additional Notation

We define the variables for the (hidden) input generating process as follows. Each $x \in \mathcal{X}$ is generated with a hidden function χ by $x = \chi(y, \xi^{(y)})$, where $\xi^{(y)} = (\xi_1^{(y)}, \dots, \xi_m^{(y)}) \in \varpi_y \subseteq \mathbb{R}^m$ is the nuisance variable. We denote the random variable for $\xi^{(y)}$ by Ξ_y; i.e., $\Xi_y(\omega_y) = \xi^{(y)}$ where $\omega_y \in \Omega_y$ is the element of the sample space Ω_y of the nuisance variable, conditioned on $Y = y$. Then, we denote the random variables for X and Z_l^s conditioned on $Y = y$ by X_y and $Z_{l,y}^s$: $X_y(\omega_y) = \chi(y, \Xi_y(\omega_y)) \in \mathcal{X}$ and $Z_{l,y}^s = \phi_l^s \circ X_y$. For any $l \in [D]$ and $y \in \mathcal{Y}$, we define the sensitivity $c_l^y(\phi_l^s)$ of the trained encoder ϕ_l^s with respect to the nuisance variable $\xi^{(y)}$ by the number such that for all $i \in [m]$, $c_l^y(\phi_l^s) \geq \sup_{\xi_1^{(y)},\dots,\xi_{i-1}^{(y)},\xi_i^{(y)},\tilde{\xi}_i^{(y)},\xi_{i+1}^{(y)},\dots,\xi_m^{(y)}} |\log p_y((\phi_l^s \circ \chi_y)(\xi_1^{(y)}, \dots, \xi_{i-1}^{(y)}, \xi_i^{(y)}, \xi_{i+1}^{(y)}, \dots, \xi_m^{(y)})) - \log p_y((\phi_l^s \circ \chi_y)(\xi_1^{(y)}, \dots, \xi_{i-1}^{(y)}, \tilde{\xi}_i^{(y)}, \xi_{i+1}^{(y)}, \dots, \xi_m^{(y)}))|$, where $\chi_y(\xi^{(y)}) = \chi(y, \xi^{(y)})$ and $p_y(q) = \mathbb{P}(Z_{l,y}^s = q)$.

For any $l \in [D+1]$ and $\lambda_l > 0$, define

$$C_{\lambda_l,l} = \frac{1}{e^{\lambda_l H(\phi_l^S)}} \sum_{q \in \mathcal{M}_l} (\mathbb{P}(\phi_l^S = q))^{1-\lambda_l}, \tag{21}$$

where $H(\phi_l^S)$ is the entropy of the random variable ϕ_l^S. We define the set of the latent variable per class by $\mathcal{Z}_{l,y}^s = \{(\phi_l^s \circ \chi_y)(\xi^{(y)}) : \xi^{(y)} \in \varpi_y\}$. For any $\gamma > 0$, we then define a (typical) subset $\mathcal{Z}_{\gamma,l,y}^s$ (of the set $\mathcal{Z}_{l,y}^s$) by $\mathcal{Z}_{\gamma,y}^{s,l} = \{z \in \mathcal{Z}_{l,y}^s : -\log \mathbb{P}(Z_{l,y}^s = z) - H(Z_{l,y}^s) \leq c_l^y(\phi_l^s)\sqrt{\frac{m \ln(\sqrt{n}/\gamma)}{2}}\}$. Let us write the element of $\mathcal{Z}_{\gamma,y}^{s,l}$ by $\mathcal{Z}_{\gamma,y}^{s,l} = \{a_1^{l,y}, \dots, a_{T_y^l}^{l,y}\}$ where $T_y^l = |\mathcal{Z}_{\gamma,y}^{s,l}|$. Finally, define maximum training loss $\mathcal{L}(f^s) = \max_{i \in \{1,\dots,n\}} \ell(f^s(x_i), y_i)$.

E.1.2. Details of Other Terms

In Theorem 1, we have that $G_1^l(q) = \frac{\mathcal{L}(f^s)\sqrt{2\gamma_l|\mathcal{Y}|}}{n^{1/4}}\sqrt{q + \ln(2|\mathcal{Y}|/\delta)} + \gamma_l \mathcal{R}(f^s)$, $\mathcal{G}_2^l = G_2^l \ln(2) + \ln(2|\mathcal{Y}|/\delta)$, $G_3^l = \max_{y \in \mathcal{Y}} \sum_{k=1}^{T_y^l} \ell(g_l^s(a_k^{l,y}), y)\sqrt{2|\mathcal{Y}|\mathbb{P}(Z_{l,y}^s = a_k^{l,y})}$, and $G_2^l = \mathbb{E}_y[c_l^y(\phi_l^s)]\sqrt{\frac{m \ln(\frac{\sqrt{n}}{\gamma_l})}{2}} + H(Z_l^s|X, Y)$. Theorem 1 holds for any (fixed) $\gamma_l > 0$.

In Theorem 2, the definitions of $G_1^l(q), G_2^l, G_3^l$ are the same as in Theorem 1, and we have that $\zeta = (I(\phi_l^S; S) + G_4^l)\ln(2) + \ln(2|\mathcal{D}|)$, $\widehat{\mathcal{G}}_2^l = \left(G_2^l + G_4^l\right)\ln(2) + \ln(4|\mathcal{Y}||\mathcal{D}|/\delta)$, $\check{\mathcal{G}}_2^l = G_4^l \ln(2) + \ln(2/\delta)$, and $G_4^l = \frac{1}{\lambda_l} \ln \frac{C_{\lambda_l,l}|\mathcal{D}|}{\delta} + H(\phi_l^S|S)$. Theorem 2 holds for any (fixed) $\gamma_l > 0$ and $\lambda_l > 0$ for all $l \in \mathcal{D}$.

Proposition 2 below shows that G_3^l can be bounded by a constant value, which is much smaller than and independent of the size of the set $\mathcal{Z}_{\gamma,l,y}^s$.

Proposition 2. *Let $l \in \{1, \dots, D\}$. Let $v_k(y) = \ell(g_l^s(a_{j_k}^{l,y}), y)^2 \mathbb{P}(Z_{l,y}^s = a_{j_k}^{l,y})$ where $k \mapsto j_k$ is a permutation of the index such that $v_1(y) \geq v_2(y) \geq \dots \geq v_{T_y^l}(y)$. If there exist some constants $\alpha_y \geq 1$ and $\beta_y, C_y > 0$ such that $v_k(y) \leq C_y e^{-(k/\beta_y)^{\alpha_y}}$, then*

$$G_3^l \leq \sqrt{2|\mathcal{Y}|} \max_{y \in \mathcal{Y}} \left(\sqrt{v_1(y)}\lceil \tilde{\beta}_y \rceil + (C_y \tilde{\beta}_y)/(\alpha_y e)\right), \tag{22}$$

without assuming that ϕ_l^s is fixed independently of the training dataset s, where $\tilde{\beta}_y = 2^{1/\alpha_y}\beta_y$.

Proof. The proof is provided in appendix F.8. □

Proposition 3 shows that the value of $\ln C_{\lambda_l,l}$ (recall from (21)) in the formula of G_4^l can be bounded by a constant value independently of $\ln|\mathcal{M}_l|$ and is much smaller than $\ln|\mathcal{M}_l|$ and $H(\phi_l^S)$:

Proposition 3. *Let $l \in \{1,\dots,D+1\}$. We denote $N = |\mathcal{M}_l|$, and enumerate $\mathcal{M}_l$ as $q_1, q_2, q_3, \cdots, q_N$ with decreasing probability, i.e. $p_i = \mathbb{P}(\phi_l^S = q_i)$ and $p_1 \geq p_2 \geq \cdots \geq p_N$.*

1. *If p_i decays sufficiently fast, i.e., $p_i \leq C/i^\alpha$ with some $\alpha > 1$ and $C \geq 1$, then for $0 < \lambda_l < 1 - 1/\alpha$, both the entropy $H(\phi_l^S)$ and $C_{\lambda_l,l}$ are bounded and independent of N:*

$$H(\phi_l^S) \leq 1 + C\alpha\left(\frac{\ln(2)}{2^\alpha} + \frac{\ln(3)}{3^\alpha} + \frac{3^{1-\alpha}((a-1)\ln(3)+1)}{(\alpha-1)^2}\right),$$
$$C_{\lambda_l,l} \leq C^{1-\lambda_l}\frac{\alpha(1-\lambda_l)}{\alpha(1-\lambda_l)-1}.$$

2. *If p_i decays slowly, i.e., $p_i = c_i/(Zi^\alpha)$ with $0 \leq \alpha < 1$ and $0 < c \leq c_i \leq C$ where Z is the normalization constant, then the entropy $H(\phi_l^S)$ grows as $\ln(N) + \mathcal{O}(1)$ where $\mathcal{O}(1)$ depends only on α, but $C_{\lambda_l,l}$ is bounded and independent of N as:*

$$C_{\lambda_l,l} \leq (\ln(1-(1-\lambda_l)\alpha) - (1-2\lambda_l)\ln(1-\alpha)) + (2-\lambda_l)\ln(C/c) + \frac{C}{c(1-\alpha)}.$$

Proof. The proof is provided in appendix F.9. □

We now discuss the factors $G_1^l(q)$ and G_2^l in Theorem 1. The formula of $G_1^l(q)$ is simplified as $G_1^l(q) = \gamma_l$ for any $q \in \mathbb{R}_{\geq 0}$ in a common scenario of deep learning where we use the 0-1 loss (to measure generalization) and have zero training error. This is because $\mathcal{L}(f^s) = 0$ and $\mathcal{R}(f^s) \leq 1$ for the scenario. In the formula of G_2, we have $H(Z_l^s|X,Y) = 0$ if the function ϕ_l^s is deterministic, which is the typical case for deep neural networks, because ϕ_l^s is the function used at inference or test time as opposed to training time (when dropout for example can be used). When the function ϕ_l^s is stochastic, we have $H(Z_l^s|X,Y) = \mathcal{O}(1)$ as $n \to \infty$. The networks can be stochastic, for example, with randomization defenses against adversarial attacks (Xie et al., 2018; Pinot et al., 2019; 2020; Levine & Feizi, 2020) or noise injections (Goldfeld et al., 2019). The value of $\mathbb{E}_y[c_l^y(\phi_l^s)]$ in the formula of G_2 measures the sensitivity with respect to the nuisance variable $\xi^{(y)}$; i.e., minimizing this value should result in better generalization, which is consistent with Theorem 1. The sensitivity $c_l^y(\phi_l^s)$ is a measure on the single final encoder ϕ_l^s; i.e., increasing the complexity of hypothesis spaces does not imply an increase in this value.

All the discussions and results, including Proposition 2, on the factors $G_1^l(q), G_2^l$, and G_3^l in Theorem 1 hold true for these factors in Theorem 2 (because we do not assume the use of the fixed encoder for these). Accordingly, we now discuss the new factor, G_4, in Theorem 2. The value of $\ln C_{\lambda_l,l}$ in the formula of G_4^l is analyzed in Proposition 3. In the formula of G_4, $H(\phi_l^S|S)$ measures the randomness of algorithm $\mathcal{A}_l$. To understand this, let us consider a real-world experiment with a coin tossing. We can model the coin tossing by a stochastic model (by saying that coin tossing has 50-50 chance of getting heads and tails) or by a deterministic model to predict the exact outcome with an exact initial condition of the physical system. similarly, we can model a single real-world algorithm with a stochastic model $\mathcal{A}_l$ (by saying that something has some random chances) or a deterministic model $\mathcal{A}_l$ with the exact initial condition, which is the random seed in the numerical experiments. That is, as in any mathematical theories and symbols, $\mathcal{A}_l$ is a theoretical placeholder with its mathematical definition; i.e., $\mathcal{A}_l$ does *not* have one-to-one correspondence to a real-world algorithm implemented in experiments. In other words, given a single real-world algorithm, there are many different ways to model the real-world algorithm and different ways result in different $\mathcal{A}_l$. For example, let us fix the real-world algorithm implemented in experiments to be one with dropout (Srivastava et al., 2014) and stochastic gradient descent (SGD). At this point, $\mathcal{A}_l$ in Theorem 2 is not fully determined yet and we can choose $\mathcal{A}_l$ differently for the same real-world algorithm by modeling the real-world differently. For instance, we can model the one with dropout and SGD as a stochastic algorithm or as a deterministic algorithm given a fixed random seed in practice. Thus, we can set $\mathcal{A}_l$ in 2 to be either a stochastic algorithm or a deterministic algorithm for the

exact same real-world and the same fixed algorithm implemented in experiments. If we set $\mathcal{A}_l$ to be a deterministic algorithm with a fixed seed, then we have $H(\phi_l^S|S) = 0$. If we set $\mathcal{A}_l$ to be a stochastic algorithm, then we increase $H(\phi_l^S|S)$ but we can potentially decrease $I(\phi_l^S; S)$ since the extra randomness can potentially reduce the mutual information of ϕ_l^S and S. Thus, there is a trade-off in how we model the real-world via $\mathcal{A}_l$ and we cannot reduce the bound arbitrarily. Our theorems allow to instantiate our bounds with both deterministic and stochastic views of the learning algorithms, without changing the real-world algorithms.

More generally, a randomized algorithm can be defined as a deterministic algorithm with an additional input that consists of a sequence of random bits (Hromkovič, 2004). Here, the sequence of random bits corresponds to the sequence of random seeds in the numerical experiments with SGD. In other words, on the one hand, we can model SGD as a stochastic process when we analyze a set of experiments with SGD over a set of random seeds that are generated randomly. On the other hand, we can model SGD as a deterministic process when we analyze one experiment with SGD for one random seed. Moreover, as Hromkovič (2004) explains, we can bridge those two cases by modeling SGD as a deterministic algorithm with its additional inputs being the seed; then (1) it is deterministic for each seed, and (2) we can recover the stochastic model by considering a sequence of the randomly generated seeds. If we analyze one experiment of SGD for one random seed, then we have a deterministic algorithm, and a typical worst-case analysis provides a guarantee on the SGD with the worst seed. But, if we analyze a set of experiments of SGD over a set of random seeds that are generated randomly, then we have a stochastic algorithm, and we can analyze its expected performance or high-probability guarantee w.r.t. the random seeds.

To formally treat the learning algorithm $\mathcal{A}_l$ as a stochastic one, we replace S with $\tilde{S}$ in eq. (5) where $\tilde{S}(\omega, \omega') = (S(\omega), \omega')$ with ω and ω' being elements of the sample spaces for S and $\mathcal{A}_l$ respectively. Similarly, when the encoder ϕ_l^s is stochastic, we replace X with $\tilde{X}$ in eq. (3) where $\tilde{X}(\omega, \omega') = (X(\omega), \omega')$ with ω and ω' being elements of the sample spaces for X and ϕ_l^s respectively. All of the proofs work in any of these cases.

E.2. On the application to the case of infinite mutual information

Section 4.2 discusses a way to apply a sample complexity bound with mutual information to the cases of infinite mutual information by using binning methods. This section considers a more general method of computing the mutual information to achieve the following goal: we demonstrate that a theoretical work on a bound with mutual information is an important and sensible research area more generally beyond our paper, even for the cases of infinite mutual information. This section also provides theoretical justifications on previous methods of computing mutual information even for the case of deterministic neural networks with continuous random variables with injective activations (Shwartz-Ziv & Tishby, 2017; Saxe et al., 2019; Chelombiev et al., 2019). We use the notation of $\phi^s = \phi_l^s$ and $g^s = g_l^s$ for a fixed l in this subsection.

This is based on the following simple observation: we can bound the generalization error of a given encoder, $\mathrm{G}[\tilde{\phi}^s]$, by using the generalization bound of another encoder, $\mathrm{B}_\delta[\phi^s]$, if we add the term measuring a distance between the two encoders, $\mathrm{D}(\phi^s, \tilde{\phi}^s)$. This is formalized in Remark 2:

Remark 2. Define $\mathrm{G}[\phi^s] = \mathbb{E}_{X,Y}[\ell_g(\phi^s(X), Y)] - \frac{1}{n}\sum_{i=1}^n \ell_g(\phi^s(x_i), y_i)$ and $\mathrm{L}[\phi^s] = \ell_g(\phi^s(X), Y)$ where $\ell_g(\phi^s(X), Y) = \ell((g^s \circ \phi^s)(X), Y)$. Suppose that for any $\delta > 0$, $\mathbb{P}(\mathrm{G}[\phi^s] \leq \mathrm{B}_\delta[\phi^s]) \geq 1 - \delta$ for some functional B_δ and that $\mathbb{P}_{X,Y}(|\mathrm{L}[\phi^s] - \mathrm{L}[\tilde{\phi}^s]| \leq \mathrm{D}(\phi^s, \tilde{\phi}^s)) = 1$ for some functional D. Then, for any $\delta > 0$, with at least probability $1 - \delta$,

$$\mathrm{G}[\tilde{\phi}^s] \leq \mathrm{B}_\delta[\phi^s] + 2\mathrm{D}(\phi^s, \tilde{\phi}^s). \tag{23}$$

Proof. Since $\mathbb{P}(|\mathrm{L}[\phi^s] - \mathrm{L}[\tilde{\phi}^s]| \leq \mathrm{D}(\phi^s, \tilde{\phi}^s)) = 1$, we have with probability one, $\mathrm{G}[\tilde{\phi}^s] \leq \mathrm{G}[\phi^s] + 2\mathrm{D}(\phi^s, \tilde{\phi}^s)$. Since $\mathbb{P}(\mathrm{G}[\phi^s] \leq \mathrm{B}[\phi^s]) \geq 1-\delta$, we have with at least probability $1-\delta$, $\mathrm{G}[\tilde{\phi}^s] \leq \mathrm{G}[\phi^s]+2\mathrm{D}(\phi^s, \tilde{\phi}^s) \leq \mathrm{B}_\delta[\phi^s]+2\mathrm{D}(\phi^s, \tilde{\phi}^s)$. □

Here, let us set $\mathrm{B}_\delta[\phi^s]$ to be a generalization bound on ϕ^s with mutual information. Then, given an original model $\tilde{\phi}^s$, its direct bound $\mathrm{B}_\delta[\tilde{\phi}^s]$ can be infinite since its mutual information can be infinite, for example, for deterministic neural networks $\tilde{\phi}^s$ with sigmoid activations for continuous random variables. However, instead of using its direct bound $\mathrm{B}_\delta[\tilde{\phi}^s]$, we can bound the generalization error of the original model $\tilde{\phi}^s$ by invoking Remark 2 to use the bound $\mathrm{B}_\delta[\phi^s]$ of another model $\phi^s \neq \tilde{\phi}^s$ such that ϕ^s has finite mutual information and $\mathrm{D}(\phi^s, \tilde{\phi}^s)$ is small.

Indeed, this is a theoretical formalization of what is implicitly done in practice when we compute mutual information of deterministic models. That is, in practice, we often compute the mutual information of the original model $\tilde{\phi}^s$ by computing

the mutual information of another model ϕ^s where ϕ^s is a binning version of $\tilde{\phi}^s$ or a noise injected version of $\tilde{\phi}^s$ with kernel density estimation (Shwartz-Ziv & Tishby, 2017; Saxe et al., 2019; Chelombiev et al., 2019).

Indeed, all of such methods of computing mutual information in experiments are theoretically valid and meaningful based on our results in Section 4.2 and Remark 2 along with Proposition 4 below, even for the case of mutual information being infinite for the original model $\tilde{\phi}^s$.

As a concrete example, we now study the case when ϕ^s is obtained from $\tilde{\phi}^s$ by injecting noise, i.e. $\phi^s(x) = \tilde{\phi}^s(x) + \lambda\vartheta$, where $\vartheta \sim \mathcal{N}(0, \mathbb{I}_d/d)$ is the Gaussian noise (d is the dimension of the intermediate output $\tilde{\phi}^s(x)$):

Proposition 4. *Let $\phi^s(x) = \tilde{\phi}^s(x) + \lambda\vartheta$, where $\vartheta \sim \mathcal{N}(0, \mathbb{I}_d/d)$. Let L be the Lipschitz constant of the function $q \mapsto \ell_g(q, Y)$ y w.r.t. the metric induced by $\|\cdot\|_2$ almost surely. Then, we can take $D(\phi^s, \tilde{\phi}^s) = \lambda L\|\vartheta\|_2$, and with probability at least $1 - 2\delta$,*

$$\mathrm{G}[\tilde{\phi}^s] \leq \mathrm{B}_\delta[\phi^s] + 2\lambda L\sqrt{\log(2/\delta)}. \tag{24}$$

Proof. Since the function $q \mapsto \ell(g_l^s(q), Y)$ is Lipschitz almost surely, we have that with probability one,

$$|\ell_g(\phi^s(X), Y) - \ell_g(\tilde{\phi}^s(X), Y)| \leq |\ell_g(\tilde{\phi}^s(X) + \lambda\vartheta, Y) - \ell_g(\tilde{\phi}^s(X), Y)| \leq \lambda L\|\vartheta\|_2.$$

Thus, we can take $D(\phi^s, \tilde{\phi}^s) = \lambda L\|\vartheta\|_2$. Since $\vartheta \sim \mathcal{N}(0, \mathbb{I}_d/d)$ is a Gaussian vector, by Bernstein inequality, $\mathbb{P}(\|\vartheta\|_2 \geq t) \leq 2e^{-t^2/2}$. If we take $t = 2\sqrt{\log(2/\delta)}$, we get $D(\phi^s, \tilde{\phi}^s) \leq \lambda L\|\vartheta\|_2 \leq 2\lambda L\sqrt{\log(2/\delta)}$ with probability $1 - \delta$. Thus, Proposition 4 follows from Remark 2 by taking union bounds. □

In Proposition 4, let us set $\mathrm{B}_\delta[\phi^s]$ to be a generalization bound on ϕ^s with mutual information $I(Z; X)$ where $Z = \phi^s \circ X$. Then, by the construction of $\phi^s(x) = \tilde{\phi}^s(x) + \lambda\vartheta$, the output is stochastic, and the mutual information $I(Z; X)$ in $\mathrm{B}_\delta[\phi^s]$ is bounded, although $I(\tilde{Z}; X)$ with $\tilde{Z} = \tilde{\phi}^s \circ X$ can be infinite. Moreover, there is a trade-off between the two terms on the righthand side of (24): injecting more noise by increasing λ reduces the mutual information $I(Z; X)$ in $\mathrm{B}_\delta[\phi^s]$, but increases error $2\lambda L\sqrt{\log(2/\delta)}$. Thus, we cannot arbitrarily change values of the bounds of the original model $\mathrm{G}[\tilde{\phi}^s]$ by choosing different methods of computing the mutual information even for the case of deterministic neural networks with continuous random variables with injective activations.

E.3. On comparisons with previous information-theoretic bounds

We discuss the difference between our bounds and the previous information-theoretic bounds (Xu & Raginsky, 2017; Bassily et al., 2018) in Section 3.2; e.g., the previous bounds do not utilize the information bottleneck term. In this subsection, we provide additional discussion on the relation between them.

We first note that Theorem 2 recovers the previous bounds if we set $\mathcal{D} = \{D + 1\}$. If $\mathcal{D} = \{D + 1\}$, then our bound in Theorem 2 removes $I(X; Z_l^s|Y)$ and only keeps $I(\phi_{D+1}^S; S) = I(f^S; S)$, resulting in the previous bounds. This is because the hypothesis space of the decoder after the output layer is always a singleton (since there is no learnable parameter) and thus there is no need of "(information) bottleneck" to avoid overfitting of such decoder. Indeed, the previous bounds only consider the setting where the hypothesis space of the decoder is a singleton; e.g., g is the identity function. In the previous bounds, since the hypothesis space of the decoder g is singleton, there is no need for the encoder to provide a bottleneck to control the complexity of the hypothesis space of the decoder. In contrast, we consider non-singleton hypothesis spaces of decoders and utilize the information bottleneck of the encoder to control the complexity of the decoder. This also illustrates the difficulty to prove our sample complexity bounds with the information bottleneck where we need to consider the non-singleton hypothesis space for the decoder.

Another challenge of proving our bounds comes from the fact that we need to efficiently utilize different sources of randomness while the previous bounds only consider the single source of randomness; i.e., $f^S = \mathcal{A}_{D+1} \circ S$ is a random variable through the randomness of training data S (and potentially of algorithm $\mathcal{A}_{D+1}$) whereas $Z_l^s = \phi_l^s \circ X$ is a random variable through the randomness of the new unseen input X (and potentially of encoder ϕ_l^s). Thus, $I(X; Z_l^s|Y)$ and $I(f^S; S)$ measures different types of mutual information with the different sources of randomness. Our bound needs to utilize both types of randomness efficiently while the previous bound only uses the randomness of S.

The main factor $I(X; Z_l^s|Y) + I(\phi_l^S; S)$ in Theorem 2 captures the novel tradeoff between the two types of mutual information. It tells us that as we minimize the information bottleneck $I(X; Z_l^s|Y)$ by optimizing ϕ_l^s based on the training

data s, we must pay the price of mutual information $I(\phi_l^S; S)$. If ϕ_l^S depends more on S, then we can more easily minimize the information bottleneck $I(X; Z_l^s|Y)$ (while minimizing the training loss for s), which comes at the cost of increasing $I(\phi_l^S; S)$. This trade-off is not captured by any of previous bounds.

As a result of utilizing the both types of randomness, we show in Section 5 that the main factor $I(X; Z_l^s|Y) + I(\phi_l^S; S)$ in our bound is a better predictor than the main factor $I(f^S; S)$ in the previous bounds.

E.4. On the standard arguments for proving the conjecture

The previous work (Shwartz-Ziv et al., 2019) provided the arguments of using the Probably Approximately Correct (PAC) bound for a finite hypothesis space $\mathcal{H}$ to obtain $\tilde{\mathcal{O}}(\sqrt{(\log|\mathcal{H}|)/n})$ (Shalev-Shwartz & Ben-David, 2014) and bounding its cardinality $|\mathcal{H}|$ via $H(Z_l^s)$. However, this argument results in the exponential factor $2^{I(X;Z_l^s)}$ as in Conjecture 1.

F. Proofs

F.1. Overview of Proofs of Theorems

Before providing complete proofs, we first provide a overview of the proofs of Theorem 1–2. Let $l \in [D]$. We first prove two properties of the typical set of Z_l, Lemma 1 and Lemma 2 (in appendix F.2), by combining a standard proof used in information theory and the McDiarmid's inequality. A typical set is a concept in information theory and we utilize the properties of a typical set to obtain the information-theoretic bounds. To achieve this, Lemma 3 (in appendix F.2) decomposes the generalization gap into four terms as $\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^n \ell(f^s(x_i), y_i) = \mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{D}$, where the one term A corresponds to the case of X being in the typical set, while other three terms B, C, and D are for the case of X being outside of the typical set. The rest of the proof of Theorem 1 analyzes each of these terms (with Lemma 1 and Lemma 2), proving that A and $\mathrm{B} + \mathrm{C} + \mathrm{D}$ are bounded by the first term and the second term on the right-hand side of eq. (6), respectively. That is, we show $\mathrm{C} + \mathrm{D} \leq \frac{\gamma_l \mathcal{R}(f^s)}{\sqrt{n}}$ in Lemma 4 (in appendix F.2) by invoking Lemma 1. Lemma 5 (in appendix F.2) then bounds the terms A and B by recasting the problem into that of multinomial distributions and by incorporating Lemma 2 into the concentration inequality of multinomial distributions.

Lemmas 1–4 (in appendix F.2) are carefully proven for the trained encoder ϕ_l^s instead of a hypothesis space of encoders ϕ_l. This is achieved by combining deterministic decompositions and probabilistic bounds with respect to the randomness of new fresh samples X instead of the training data S. In contrast, Lemma 5 (in appendix F.2) is proven for a hypothesis space Φ of encoders using the randomness of S, where Φ must be independent of s. These decompositions and probabilistic bounds for different sample spaces enable the exponential improvement over the previous bounds. Combining Lemmas 3-5 (in appendix F.2) produces Lemma 6, which proves Theorem 1 by setting the hypothesis space as $\Phi = \{\phi_l^s\}$ where ϕ_l^s is fixed independently of s.

The standard proof techniques result in the exponential factor $2^{I(X;Z_l^s)}$ as in Conjecture 1 (see appendix E for more details). This paper provides a novel proof technique to avoid the exponential factor. Compared to arguments for Conjecture 1, our proof discards the non-mathematical arguments regarding the typical set, keeps track of all the effects of the approximation and non-typicality rigorously, and discards the assumption of the input dimension approaching infinity with an ergodic Markov random field.

Another main challenge in proving our main result, Theorem 2, is avoiding the dependence on the hypothesis space for the value of $I(X; Z_l^s|Y)$. That is, with a relatively simpler proof, we could prove a similar bound with $\sup_{\phi_l \in \Phi_l} I(X; \phi_l \circ X|Y)$ where Φ_l is a fixed hypothesis space of the encoder ϕ_l. However, this dependence on the hypothesis space is not preferred since enlarging the hypothesis space can increase the value, whereas the value of $I(X; Z_l^s)$ in our bound is independent of the hypothesis space given the final hypothesis ϕ_l^s.

We carefully construct and prove our key lemmas in the following subsection, which enables us to avoid the dependence over the entire hypothesis space and the exponential factor.

F.2. Proofs of Key Lemmas

We use the notation of $\ln = \log_e$ and $\log = \log_2$. Fix $l \in \{1, \ldots, D\}$ throughout this section. For the simplicity of the notation, we write $Z = Z_l^s$ and $Z_y = Z_{l,y}^s$ in the following; we must to be always aware of the dependence on s for related

variables. We recall that

$$Z_y = \phi_l^s \circ X_y.$$

We write $\xi^{(y)} \in \varpi_y \subseteq \mathbb{R}^m$, and define the set of the latent variable per class by

$$\mathcal{Z}_y = \left\{ (\phi_l^s \circ \chi_y)(\xi^{(y)}) : \ \xi^{(y)} \in \varpi_y \right\}.$$

For any $\gamma > 0$, we then define the typical subset $\mathcal{Z}_{\gamma,y}^s$ of the set $\mathcal{Z}_y$ by

$$\mathcal{Z}_{\gamma,y}^s = \left\{ z \in \mathcal{Z}_y : -\log \mathbb{P}(Z_y = z) - H(Z_y) \leq c_l^y(\phi_l^s) \sqrt{\frac{m \ln(\sqrt{n}/\gamma)}{2}} \right\}.$$

Then, for any set A and any function φ, we have that

$$\begin{aligned}
\mathbb{P}(Z \in A | Y = y) = \mathbb{P}(Z_y \in A) &= \mathbb{P}(\{\omega_y \in \Omega_y : Z_y(\omega_y) \in A\}) \\
&= \mathbb{P}(\{\omega_y \in \Omega_y : (\phi_l^s \circ \chi_y)(\Xi_y(\omega_y)) \in A\}) \\
&= \mathbb{P}((\phi_l^s \circ \chi_y \circ \Xi_y) \in A),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}((\varphi \circ Z) > 0 | Y = y) = \mathbb{P}((\varphi \circ Z_y) > 0) &= \mathbb{P}(\{\omega_y \in \Omega_y : \varphi(Z_y(\omega_y)) > 0\}) \\
&= \mathbb{P}(\{\omega_y \in \Omega_y : \varphi((\phi_l^s \circ \chi_y)(\Xi_y(\omega_y))) > 0\}) \\
&= \mathbb{P}((\varphi \circ \phi_l^s \circ \chi_y \circ \Xi_y) > 0).
\end{aligned}$$

Thus, for example, we can write

$$\begin{aligned}
\mathbb{P}(Z \notin \mathcal{Z}_{\gamma,y}^s | Y = y) &= \mathbb{P}\left((\phi_l^s \circ \chi_y \circ \Xi_y) \notin \mathcal{Z}_{\gamma,y}^s\right) \\
&= \mathbb{P}\left(\{\omega_y \in \Omega_y : -\log \mathbb{P}(Z_y = \phi_l^s(X_y(\omega_y))) - H(Z_y) > \epsilon\}\right) \\
&= \mathbb{P}\left(\left\{\omega_y \in \Omega_y : -\log \mathbb{P}\left(\{\omega_y' \in \Omega_y : Z_y(\omega_y') = Z_y(\omega_y)\}\right) - H(Z_y) > \epsilon\right\}\right),
\end{aligned}$$

where $\epsilon = c_l^y(\phi_l^s)\sqrt{\frac{m \ln(\sqrt{n}/\gamma)}{2}}$.

F.2.1. Probability of Going outside of The Typical Subset

The following lemma shows that the conditional probability of going outside of $\mathcal{Z}_{\gamma,y}$ is bounded by $\frac{\gamma}{\sqrt{n}}$:

Lemma 1. *For any $\gamma > 0$, it holds that*

$$\mathbb{P}(Z \notin \mathcal{Z}_{\gamma,y}^s \mid Y = y) \leq \frac{\gamma}{\sqrt{n}}.$$

Proof. Fix $y \in \mathcal{Y}$. We then write $\xi = \xi^{(y)}$ for the simplicity of the notation. We now consider the statistical property of the function $\xi \mapsto -\log \mathbb{P}(Z_y = \phi_l^s(\chi(y, \xi)))$. That is, in the following, we will apply McDiarmid's inequality w.r.t. the sample space $\omega_y \in \Omega_y$ to the following function:

$$\xi \mapsto -\log \mathbb{P}(Z_y = \phi_l^s(\chi(y, \xi))) = -\log \mathbb{P}(\{\omega_y' \in \Omega_y : Z_y(\omega_y') = \phi_l^s(\chi(y, \xi))\}).$$

For the simpler notation, define the function p_y by

$$p_y(q) = \mathbb{P}(Z_y = q).$$

Then, we can rewrite the above function of ξ as

$$\xi = (\xi_1, \ldots, \xi_m) \mapsto -\log p_y(\phi_l^s(\chi(y, \xi))).$$

We also define

$$\tilde{\mathcal{Z}}_{\epsilon,y}^s = \{z \in \mathcal{Z}_y : -\log p_y(z) - H(Z) \leq \epsilon\}.$$

For any (h, φ) and $t = h(q)$ with a probability mass function p, since $p(t) = \sum_{q \in h^{-1}(t)} p(q)$,

$$\begin{aligned}\mathbb{E}_{t\sim p}\left[\varphi(t)\right] = \sum_t \varphi(t)p(t) &= \sum_t \varphi(t) \sum_{q\in h^{-1}(t)} p(q) \\ &= \sum_t \sum_{q\in h^{-1}(t)} \varphi(t)p(q) \\ &= \sum_t \sum_{q\in h^{-1}(t)} \varphi(h(q))p(q) \\ &= \sum_q \varphi(h(q))p(q) = \mathbb{E}_{q\sim p}[\varphi(h(q))].\end{aligned}$$

Thus, by choosing $q = \xi_y$, $h(q) = \phi_l^s(\chi(y, q))$, and $\varphi(t) = -\log p_y(t)$, we have that

$$\mathbb{E}_{\Xi_y}\left[-\log p_y(\phi_l^s(\chi(y, \Xi_y)))\right] = \mathbb{E}_q[\varphi(h(q))] = \mathbb{E}_t\left[\varphi(t)\right] = \mathbb{E}_{Z_y}\left[-\log p_y(Z_y)\right] = H(Z_y).$$

Thus, by using McDiarmid's inequality,

$$\mathbb{P}_{\Xi_y}\left(-\log p_y((\phi_l^s \circ \chi_y)(\Xi_y)) - H(Z_y) \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{mc_l^y(\phi_l^s)^2}\right).$$

By setting $\delta = \exp\left(-\frac{2\epsilon^2}{c_l^y(\phi_l^s)^2}\right)$ and solving for ϵ, we set

$$\epsilon = c_l^y(\phi_l^s)\sqrt{\frac{m\ln(1/\delta)}{2}},$$

with which

$$\begin{aligned}\mathbb{P}(Z \notin \tilde{\mathcal{Z}}_{\epsilon,y}^s \mid Y = y) &= \mathbb{P}_{\Xi_y}\left((\phi_l^s \circ \chi_y)(\Xi_y) \notin \tilde{\mathcal{Z}}_{\epsilon,y}^s\right) \\ &= \mathbb{P}_{\Xi_y}\left(-\log p_y((\phi_l^s \circ \chi_y)(\Xi_y)) - H(Z_y) > \epsilon\right) \\ &\leq \mathbb{P}_{\Xi_y}\left(-\log p_y((\phi_l^s \circ \chi_y)(\Xi_y)) - H(Z_y) \geq \epsilon\right) \leq \delta.\end{aligned}$$

Therefore, by setting $\delta = \frac{\gamma}{\sqrt{n}}$ and accordingly $\epsilon = c_l^y(\phi_l^s)\sqrt{\frac{m\ln(1/\delta)}{2}} = c_l^y(\phi_l^s)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}$, we have proven the desired statement, since $\tilde{\mathcal{Z}}_{\epsilon,y}^s = \mathcal{Z}_{\gamma,y}^s$ when $\epsilon = c_l^y(\phi_l^s)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}$. □

F.2.2. Size of the Typical Subset

The following lemmas bounds the size of the subset $\mathcal{Z}_{\gamma,y}^s$:

Lemma 2. *For any $\gamma > 0$,*

$$|\mathcal{Z}_{\gamma,y}^s| \leq 2^{H_y(Z_y) + c_l^y(\phi_l^s)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}}.$$

Proof. Set $\epsilon = c_l^y(\phi_l^s)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}$. We define the function p_y by $p_y(q) = \mathbb{P}(Z_y = q)$. Then, from the definition of $\mathcal{Z}_{\gamma,y}^s$, we have that for any $a \in \mathcal{Z}_{\gamma,y}^s$,

$$\begin{aligned}-\log p_y(a) - H(Z_y) \leq \epsilon &\iff -\log p_y(a) \leq H(Z_y) + \epsilon \\ &\iff -\left(H(Z_y) + \epsilon\right) \leq \log p_y(a) \\ &\iff 2^{-H(Z_y)-\epsilon} \leq p_y(a).\end{aligned}$$

Using $2^{-H(Z_y)-\epsilon} \leq p_y(a) = \mathbb{P}(Z_y = a)$ for all $a \in \mathcal{Z}_{\gamma,y}^s$,

$$1 \geq \mathbb{P}(Z_y \in \mathcal{Z}_{\gamma,y}^s) = \sum_{a\in\mathcal{Z}_{\gamma,y}^s} \mathbb{P}(Z_y = a) \geq \sum_{a\in\mathcal{Z}_{\gamma,y}^s} 2^{-H(Z_y)-\epsilon} = |\mathcal{Z}_{\gamma,y}^s| 2^{-H(Z_y)-\epsilon}.$$

This implies that using $\epsilon = c_l^y(\phi_l^s)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}$,

$$|\mathcal{Z}_{\gamma,y}^s| \leq 2^{H(Z_y)+\epsilon} = 2^{H(Z_y)+c_l^y(\phi_l^s)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}}.$$

□

F.2.3. Decomposition of Expected Loss using the Typical Subset

Let us write

$$z_i = \phi_l^s(x_i) \in \mathcal{Z}_l \subseteq \mathbb{R}^{m_l},$$

and

$$\ell_l(q, y) = \ell(g_l^s(q), y).$$

Then, by the law of the unconscious statistician,

$$\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^n \ell(f^s(x_i), y_i) = \mathbb{E}_{Z,Y}[\ell_l(Z, Y)] - \frac{1}{n}\sum_{i=1}^n \ell_l(z_i, y_i).$$

For simplicity of the notation, define $A_y = \mathcal{Z}_{\gamma,y}^s$. We now consider a partition of the space $\mathcal{Z}_l$ as $\mathcal{Z}_l = \{z \in A_y\} \cup \{z \notin A_y\}$. Fix an order and write the element of A_y by $A_y = \{a_1^y, \ldots, a_{T_y}^y\}$ where $T_y = |A_y| \leq 2^{H_y(\phi_l \circ X_y)+c_l^y(\phi_l^s)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}}$ from the Lemma 2. We define $\mathcal{I}_y = \{i \in [n] : y_i = y\}$, $\tilde{\mathcal{I}}^y = \{i \in [n] : z_i \notin A_y, y_i = y\}$, $\mathcal{I}_k^y = \{i \in [n] : z_i = a_k^y, y_i = y\}$, $\tilde{\mathcal{Y}} = \{y \in \mathcal{Y} : |\tilde{\mathcal{I}}^y| \neq 0\}$, $\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y} \ell_l(z_i, y)q \triangleq 0$ for any q if $|\tilde{\mathcal{I}}^y| = 0$, and $\frac{1}{|\mathcal{I}_k^y|}\sum_{i\in\mathcal{I}_k^y} \ell_l(z_i, y)q \triangleq 0$ for any q if $|\mathcal{I}_k^y| = 0$. Here, for example, Z, a_k^y, A_y, $|\mathcal{I}_k^y|$, and $|\tilde{\mathcal{I}}^y|$ depend on the training dataset s through the function ϕ_l^s due to their definitions.

Using these, we can decompose the expected loss as in the following lemma:

Lemma 3. *The following holds (deterministically):*

$$\begin{aligned}\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^n \ell(f^s(x_i), y_i) = &\sum_{y\in\tilde{\mathcal{Y}}} \frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y} \ell_l(z_i, y)\left(\mathbb{P}(Y = y, Z \notin A_y) - \frac{|\tilde{\mathcal{I}}^y|}{n}\right) \\ &+ \sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y} \ell_l(a_k^y, y)\left(\mathbb{P}(Y = y, Z = a_k^y) - \frac{|\mathcal{I}_k^y|}{n}\right) \\ &+ \sum_{y\in\mathcal{Y}} \mathbb{P}(Y = y, Z \notin A_y)\mathbb{E}_{Z,Y}[\ell_l(Z, Y)|Z \notin A_y, Y = y] \\ &- \sum_{y\in\tilde{\mathcal{Y}}} \mathbb{P}(Y = y, Z \notin A_y)\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y} \ell_l(z_i, y).\end{aligned} \tag{25}$$

Proof. we can decompose the expected loss by using conditionals as

$$\mathbb{E}_{Z,Y}[\ell_l(Z, Y)] = \sum_{y\in\mathcal{Y}} \mathbb{P}(Y = y)\mathbb{E}_{Z,Y}[\ell_l(Z, Y)|Y = y].$$

Furthermore, we can decompose the conditional expectation as

$$\begin{aligned}\mathbb{E}_{Z,Y}[\ell_l(Z, Y)|Y = y] &= \mathbb{P}(Z \notin A_y|Y = y)\mathbb{E}_{Z,Y}[\ell_l(Z, Y)|Z \notin A_y, Y = y] \\ &\quad + \mathbb{P}(Z \in A_y|Y = y)\mathbb{E}_{Z,Y}[\ell_l(Z, Y)|Z \in A_y, Y = y] \\ &= \mathbb{P}(Z \notin A_y|Y = y)\mathbb{E}_{Z,Y}[\ell_l(Z, Y)|Z \notin A_y, Y = y] \\ &\quad + \sum_{k=1}^{T_y} \mathbb{P}(Z = a_k^y|Y = y)\mathbb{E}_{Z,Y}[\ell_l(Z, Y)|Z = a_k^y, Y = y]\end{aligned}$$

$$= \mathbb{P}(Z \notin A_y | Y = y) \mathbb{E}_{Z,Y}[\ell_l(Z,Y) | Z \notin A_y, Y = y] + \sum_{k=1}^{T_y} \mathbb{P}(Z = a_k^y | Y = y) \ell_l(a_k^y, y)$$

Summarising above,

$$\mathbb{E}_{Z,Y}[\ell_l(Z,Y)] = \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y, Z \notin A_y) \mathbb{E}_{Z,Y}[\ell_l(Z,Y) | Z \notin A_y, Y = y] + \sum_{y \in \mathcal{Y}} \sum_{k=1}^{T_y} \mathbb{P}(Y = y, Z = a_k^y) \ell_l(a_k^y, y).$$

Similarly, we can decompose the training loss as

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \ell_l(z_i, y_i) &= \frac{1}{n} \sum_{y \in \mathcal{Y}} \sum_{i \in \mathcal{I}_y} \ell_l(z_i, y) \\
&= \frac{1}{n} \sum_{y \in \mathcal{Y}} \left(\sum_{i \in \tilde{\mathcal{I}}^y} \ell_l(z_i, y) + \sum_{k=1}^{T_y} \sum_{i \in \mathcal{I}_k^y} \ell_l(z_i, y) \right) \\
&= \sum_{y \in \tilde{\mathcal{Y}}} \frac{|\tilde{\mathcal{I}}^y|}{n} \frac{1}{|\tilde{\mathcal{I}}^y|} \sum_{i \in \tilde{\mathcal{I}}^y} \ell_l(z_i, y) + \sum_{y \in \mathcal{Y}} \sum_{k=1}^{T_y} \frac{|\mathcal{I}_k^y|}{n} \ell_l(a_k, y)
\end{aligned}$$

Using these, we now decompose the expected loss as follows:

$$\begin{aligned}
\mathbb{E}_{Z,Y}[\ell_l(Z,Y)] - \frac{1}{n} \sum_{i=1}^{n} \ell_l(z_i, y_i) &= \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y, Z \notin A_y) \mathbb{E}_{Z,Y}[\ell_l(Z,Y) | Z \notin A_y, Y = y] \\
&\quad + \sum_{y \in \mathcal{Y}} \sum_{k=1}^{T_y} \mathbb{P}(Y = y, Z = a_k^y) \ell_l(a_k^y, y) - \frac{1}{n} \sum_{i=1}^{n} \ell_l(z_i, y_i) \\
&\quad \pm \sum_{y \in \tilde{\mathcal{Y}}} \mathbb{P}(Y = y, Z \notin A_y) \frac{1}{|\tilde{\mathcal{I}}^y|} \sum_{i \in \tilde{\mathcal{I}}^y} \ell_l(z_i, y) \pm \sum_{y \in \mathcal{Y}} \sum_{k=1}^{T_y} \frac{|\mathcal{I}_k^y|}{n} \ell_l(a_k, y)
\end{aligned}$$

By rearranging,

$$\begin{aligned}
&\mathbb{E}_{Z,Y}[\ell_l(Z,Y)] - \frac{1}{n} \sum_{i=1}^{n} \ell_l(z_i, y_i) \\
&= \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y, Z \notin A_y) \mathbb{E}_{Z,Y}[\ell_l(Z,Y) | Z \notin A_y, Y = y] \\
&\quad - \sum_{y \in \tilde{\mathcal{Y}}} \mathbb{P}(Y = y, Z \notin A_y) \left(\frac{1}{|\tilde{\mathcal{I}}^y|} \sum_{i \in \tilde{\mathcal{I}}^y} \ell_l(z_i, y) \right) \\
&\quad + \sum_{y \in \mathcal{Y}} \sum_{k=1}^{T_y} \ell_l(a_k^y, y) \left(\mathbb{P}(Y = y, Z = a_k^y) - \frac{|\mathcal{I}_k^y|}{n} \right) \\
&\quad + \sum_{y \in \tilde{\mathcal{Y}}} \mathbb{P}(Y = y, Z \notin A_y) \frac{1}{|\tilde{\mathcal{I}}^y|} \sum_{i \in \tilde{\mathcal{I}}^y} \ell_l(z_i, y) + \sum_{y \in \mathcal{Y}} \sum_{k=1}^{T_y} \frac{|\mathcal{I}_k^y|}{n} \ell_l(a_k, y) - \frac{1}{n} \sum_{i=1}^{n} \ell_l(z_i, y_i) \\
&= \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y, Z \notin A_y) \mathbb{E}_{Z,Y}[\ell_l(Z,Y) | Z \notin A_y, Y = y]
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{y\in\tilde{\mathcal{Y}}} \mathbb{P}(Y=y, Z\notin A_y)\left(\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)\right)\\
&+ \sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\ell_l(a_k^y,y)\left(\mathbb{P}(Y=y,Z=a_k^y)-\frac{|\mathcal{I}_k^y|}{n}\right)\\
&+ \sum_{y\in\tilde{\mathcal{Y}}} \mathbb{P}(Y=y, Z\notin A_y)\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y) + \sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\frac{|\mathcal{I}_k^y|}{n}\ell_l(a_k,y)\\
&- \sum_{y\in\tilde{\mathcal{Y}}}\frac{|\tilde{\mathcal{I}}^y|}{n}\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y) - \sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\frac{|\mathcal{I}_k^y|}{n}\ell_l(a_k,y)
\end{aligned}
$$

By combining the relevant terms,

$$
\begin{aligned}
&\mathbb{E}_{Z,Y}[\ell_l(Z,Y)] - \frac{1}{n}\sum_{i=1}^{n}\ell_l(z_i,y_i)\\
&= \sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y,Z\notin A_y)\mathbb{E}_{Z,Y}[\ell_l(Z,Y)|Z\notin A_y, Y=y]\\
&\quad - \sum_{y\in\tilde{\mathcal{Y}}} \mathbb{P}(Y=y, Z\notin A_y)\left(\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)\right)\\
&\quad + \sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\ell_l(a_k^y,y)\left(\mathbb{P}(Y=y,Z=a_k^y)-\frac{|\mathcal{I}_k^y|}{n}\right)\\
&\quad + \sum_{y\in\tilde{\mathcal{Y}}}\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)\left(\mathbb{P}(Y=y,Z\notin A_y)-\frac{|\tilde{\mathcal{I}}^y|}{n}\right) + \sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\frac{|\mathcal{I}_k^y|}{n}\left(\ell_l(a_k,y)-\ell_l(a_k,y)\right)\\
&= \sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y,Z\notin A_y)\mathbb{E}_{Z,Y}[\ell_l(Z,Y)|Z\notin A_y, Y=y]\\
&\quad - \sum_{y\in\tilde{\mathcal{Y}}} \mathbb{P}(Y=y, Z\notin A_y)\left(\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)\right)\\
&\quad + \sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\ell_l(a_k^y,y)\left(\mathbb{P}(Y=y,Z=a_k^y)-\frac{|\mathcal{I}_k^y|}{n}\right)\\
&\quad + \sum_{y\in\tilde{\mathcal{Y}}}\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)\left(\mathbb{P}(Y=y,Z\notin A_y)-\frac{|\tilde{\mathcal{I}}^y|}{n}\right)
\end{aligned}
$$

This implies the desired statement. □

F.2.4. Bounding the third and forth terms in the decomposition

Define

$$R_y = \mathbb{E}_{Z,Y}[\ell_l(Z,Y)|Z\notin A_y, Y=y].$$

Then, the following lemma bounds the third and forth terms in the decomposition of (25) from the previous subsection:

Lemma 4. *For any $\gamma > 0$, the following holds:*

$$\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)\frac{\gamma R_y}{\sqrt{n}} \geq \sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y,Z\notin A_y)\mathbb{E}_{Z,Y}[\ell_l(Z,Y)|Z\notin A_y, Y=y]$$

$$-\sum_{y\in\tilde{\mathcal{Y}}}\mathbb{P}(Y=y,Z\notin A_y)\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y). \tag{26}$$

Proof. Recalling the definition of $A_y=\mathcal{Z}^s_{\gamma,y}$, the third term can be written as

$$\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y,Z\notin A_y)R_y=\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)\mathbb{P}(Z\notin\mathcal{Z}^s_{\gamma,y}|Y=y)R_y.$$

Then, using Lemma 1, for any $\gamma>0$,

$$\mathbb{P}(Z\notin\mathcal{Z}^s_{\gamma,y}|Y=y)\le\frac{\gamma}{\sqrt{n}}.$$

Since $\sum_{y\in\tilde{\mathcal{Y}}}\mathbb{P}(Y=y,Z\notin A_y)\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)\ge 0$, combining these implies the desired statement. □

F.2.5. Bounding the first and second term in the decomposition

Let Φ be fixed such that Φ is independent of s, while Φ can depend on the underlying data distribution. The following lemma probabilistically bounds the first and second term in the decomposition of (25):

Lemma 5. *If $\phi_l^s\in\Phi$, for any $\gamma>0$ and $\delta>0$, with probability at least $1-\delta$, the following holds:*

$$\begin{aligned}
&\sum_{y\in\tilde{\mathcal{Y}}}\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)\left(\mathbb{P}(Y=y,Z\notin A_y)-\frac{|\tilde{\mathcal{I}}^y|}{n}\right)\\
&\le\left(\sum_{y\in\tilde{\mathcal{Y}}}\sqrt{\mathbb{P}(Z\notin\mathcal{Z}^s_{\gamma,y},Y=y)}\frac{\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)}{|\tilde{\mathcal{I}}^y|}\right)\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}},\ \textit{and,}\\
&\sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\left(\mathbb{P}(Y=y,Z=a^y_k)-\frac{|\mathcal{I}^y_k|}{n}\right)\\
&\le\sum_{y\in\mathcal{Y}}\left(\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\sqrt{\mathbb{P}(Z=a^y_k,Y=y)}\right)\sqrt{\frac{2\left(I(X_y;Z_y)+G^y_2\right)\ln(2)+2\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}.
\end{aligned}\tag{27}$$

where

$$G^y_2=c^y_l(\phi^s_l)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}+H(Z_y|X_y).$$

Proof. Let $\gamma>0$ fixed. Define

$$\mathcal{X}_y=\left\{\chi_y(\xi^{(y)}):\ \xi^{(y)}\in\varpi_y\right\},$$

and

$$\hat{A}_y(\phi_l)=\left\{x\in\mathcal{X}_y:-\log\mathbb{P}(Z_y=\phi_l(x))-H(Z_y)\le c^y_l(\phi_l)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}\right\}.$$

For each ϕ_l, write the element of $\hat{A}_y(\phi_l)$ by $\hat{A}_y(\phi_l)=\{\hat{a}^y_1(\phi_l),\dots,\hat{a}^y_{\hat{T}_y(\phi_l)}(\phi_l)\}$ (with a fixed order) where $\hat{T}_y(\phi_l)=|\hat{A}_y(\phi_l)|$. Moreover, we define

$$\hat{\mathcal{I}}^y_k(\phi_l)=\begin{cases}\{i\in[n]:\phi_l(x_i)=\hat{a}^y_k(\phi_l),y_i=y\} & \text{if } k\in[\hat{T}_y(\phi_l)]\\ \{i\in[n]:\phi_l(x_i)\notin\hat{A}_y(\phi_l),y_i=y\} & \text{if } k=\hat{T}_y(\phi_l)+1.\end{cases}$$

These are defined such that the previously defined notations are recovered when we set $\phi_l=\phi^s_l$ as

$$\mathcal{Z}^s_{\gamma,y}=\hat{A}_y(\phi^s_l),\quad A_y=\hat{A}_y(\phi^s_l), \tag{28}$$

$$(a_1^y, \dots, a_T^y) = (\hat{a}_1^y(\phi_l^s), \dots, \hat{a}^y_{\hat{T}_y(\phi_l^s)}(\phi_l^s)), \quad \mathcal{I}_k^y = \hat{\mathcal{I}}_k^y(\phi_l^s),$$
$$\tilde{\mathcal{I}}^y = \hat{\mathcal{I}}^y_{\hat{T}_y(\phi_l^s)+1}(\phi_l^s), \quad T_y = \hat{T}_y(\phi_l^s).$$

We begin with bounding terms for a fixed encoder, before extending it to the case of encoders learned from the training set. Let ϕ_l fixed and define

$$p_k^y = \begin{cases} \mathbb{P}((\phi_l \circ X) = \hat{a}_k^y(\phi_l), Y = y) & \text{if } k \in [\hat{T}_y(\phi_l)] \\ \mathbb{P}((\phi_l \circ X) \notin \hat{A}_y(\phi_l), Y = y) & \text{if } k = \hat{T}_y(\phi_l) + 1. \end{cases}$$

Let $y \in \mathcal{Y}$ and $k \in [\hat{T}_y(\phi_l) + 1]$. Then, we first prove the following statement: for any $\delta > 0$, with probability at least $1 - \delta$,

$$p_k^y - \frac{|\hat{\mathcal{I}}_k^y(\phi_l)|}{n} \leq \sqrt{\frac{2p_k^y \ln(1/\delta)}{n}}. \tag{29}$$

To prove this statement, fix $y \in \mathcal{Y}$ and $k \in [\hat{T}_y(\phi_l) + 1]$. Let us write $\hat{\mathcal{I}}_k = \hat{\mathcal{I}}_k^y(\phi_l)$ and $p_k = p_k^y$. If $p_k = 0$, then the desired statement holds trivially because $p_k - \frac{|\hat{\mathcal{I}}_k|}{n} = -\frac{|\hat{\mathcal{I}}_k|}{n} \leq \sqrt{\frac{2p_k \ln(1/\delta)}{n}}$ where $\frac{|\hat{\mathcal{I}}_k|}{n} = 0$ and $\sqrt{\frac{2p_i \ln(1/\delta)}{n}} = 0$. Thus, for the rest, we consider the case where $p_k \neq 0$. We notice that $(|\hat{\mathcal{I}}_1|, \dots, |\hat{\mathcal{I}}_{T+1}|)$ follows the multinomial distribution with parameter n and $(p_1, \dots, p_{T+1})$. Thus, we invoke Lemma 3 of (Kawaguchi et al., 2022a) with $\bar{a}_i = 1$ and $\bar{a}_j = 0$ for all $j \neq i$ (which satisfies $\sum_{i=1}^K \bar{a}_i p_i \neq 0$ since $p_i \neq 0$), yielding that for any $M > 0$,

$$\mathbb{P}\left(p_k - \frac{|\hat{\mathcal{I}}_k|}{n} > M\right) \leq \exp\left(-\frac{nM^2}{2p_k}\right).$$

By setting $M = \sqrt{\frac{2p_i \ln(1/\delta)}{n}}$,

$$\mathbb{P}\left(p_k - \frac{|\hat{\mathcal{I}}_k|}{n} > \sqrt{\frac{2p_k \ln(1/\delta)}{n}}\right) \leq \delta.$$

This proves the statement of (29). Using (29), we can bound the first and second terms for a fixed ϕ_l as follows. For the first term with a fixed ϕ_l, using (29), by taking union bounds over all $y \in \mathcal{Y}$, we have that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $y \in \mathcal{Y}$:

$$\begin{aligned} &\mathbb{P}(\phi_l(X) \notin \hat{A}_y(\phi_l), Y = y) - \frac{|\hat{\mathcal{I}}^y_{\hat{T}_y(\phi_l)+1}(\phi_l)|}{n} \\ &\leq \sqrt{\frac{2\mathbb{P}(\phi_l(X) \notin \hat{A}_y(\phi_l), Y = y) \ln(|\mathcal{Y}|/\delta)}{n}}. \end{aligned} \tag{30}$$

For the second term with a fixed ϕ_l, using (29), by taking union bounds over all $y \in \mathcal{Y}$ and all $k \in [\hat{T}_y(\phi_l)]$, we have that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $y \in \mathcal{Y}$ and all $k \in [\hat{T}_y(\phi_l)]$,

$$\begin{aligned} &\mathbb{P}(\phi_l(X) = \hat{a}_k^y(\phi_l), Y = y) - \frac{|\hat{\mathcal{I}}_k^y(\phi_l)|}{n} \\ &\leq \sqrt{\mathbb{P}(\phi_l(X) = \hat{a}_k^y(\phi_l), Y = y)} \sqrt{\frac{2\ln(|\mathcal{Y}|\hat{T}_y(\phi_l)/\delta)}{n}}. \end{aligned}$$

We now extend the results for the case of encoders learned from the training set; i.e., ϕ_l is no longer fixed. By taking union bounds with the previous two bounds, we have that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $\phi_l \in \Phi$:

$$\mathbb{P}(\phi_l(X) \notin \hat{A}_y(\phi_l), Y = y) - \frac{|\hat{\mathcal{I}}^y_{\hat{T}_y(\phi_l)+1}(\phi_l)|}{n}$$

$$\leq \sqrt{\frac{2\mathbb{P}(\phi_l(X) \notin \hat{A}_y(\phi_l), Y=y)\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}},$$

and for all $k \in [\hat{T}_y(\phi_l)]$,

$$\mathbb{P}(\phi_l(X) = \hat{a}_k^y(\phi_l), Y=y) - \frac{|\hat{\mathcal{I}}_k^y(\phi_l)|}{n}$$
$$\leq \sqrt{\mathbb{P}(\phi_l(X) = \hat{a}_k^y(\phi_l), Y=y)}\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|\hat{T}_y(\phi_l)/\delta)}{n}}.$$

Thus, if $\phi_l^s \in \Phi$, then we have that for any $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$\mathbb{P}(\phi_l^s(X) \notin \hat{A}_y(\phi_l^s), Y=y) - \frac{|\hat{\mathcal{I}}^y_{\hat{T}_y(\phi_l^s)+1}(\phi_l^s)|}{n}$$
$$\leq \sqrt{\frac{2\mathbb{P}(\phi_l^s(X) \notin \hat{A}_y(\phi_l^s), Y=y)\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}.$$

and for all $k \in [\hat{T}_y(\phi_l^s)]$,

$$\mathbb{P}(\phi_l^s(X) = \hat{a}_k^y(\phi_l^s), Y=y) - \frac{|\hat{\mathcal{I}}_k^y(\phi_l^s)|}{n}$$
$$\leq \sqrt{\mathbb{P}(\phi_l^s(X) = \hat{a}_k^y(\phi_l^s), Y=y)}\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|\hat{T}_y(\phi_l^s)/\delta)}{n}}.$$

By using (28), this means that if $\phi_l^s \in \Phi$, for any $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$\mathbb{P}(Z \notin \mathcal{Z}^s_{\gamma,y}, Y=y) - \frac{|\tilde{\mathcal{I}}^y|}{n} \leq \sqrt{\frac{2\mathbb{P}(Z \notin \mathcal{Z}^s_{\gamma,y}, Y=y)\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}},$$

and for all $k \in [T_y]$,

$$\mathbb{P}(Z = a_k^y, Y=y) - \frac{|\mathcal{I}_k^y|}{n} \leq \sqrt{\mathbb{P}(Z = a_k^y, Y=y)}\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|T_y/\delta)}{n}}.$$

Since $\ell_l(z_i, y) \geq 0$ and $\ell_l(a_k^y, y) \geq 0$, this implies that if $\phi_l^s \in \Phi$, for any $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$\sum_{y\in\tilde{\mathcal{Y}}} \frac{1}{|\tilde{\mathcal{I}}^y|} \sum_{i\in\tilde{\mathcal{I}}^y} \ell_l(z_i, y)\left(\mathbb{P}(Y=y, Z \notin A_y) - \frac{|\tilde{\mathcal{I}}^y|}{n}\right)$$
$$\leq \left(\sum_{y\in\tilde{\mathcal{Y}}} \sqrt{\mathbb{P}(Z \notin \mathcal{Z}^s_{\gamma,y}, Y=y)}\frac{\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)}{|\tilde{\mathcal{I}}^y|}\right)\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}},$$

and for all $k \in [T_y]$,

$$\sum_{y\in\mathcal{Y}}\sum_{k=1}^{T_y} \ell_l(a_k^y, y)\left(\mathbb{P}(Y=y, Z=a_k^y) - \frac{|\mathcal{I}_k^y|}{n}\right)$$
$$\leq \sum_{y\in\mathcal{Y}}\left(\sum_{k=1}^{T_y}\ell_l(a_k^y,y)\sqrt{\mathbb{P}(Z=a_k^y, Y=y)}\right)\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|T_y/\delta)}{n}}.$$

Here, using Lemma 2, we have that $T_y = |\mathcal{Z}^s_{\gamma,y}| \leq 2^{H(Z_y)+c^y_l(\phi^s_l)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}}$. Thus,

$$\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|T_y/\delta)}{n}} = \sqrt{\frac{2\ln(T_y)+2\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}$$
$$\leq \sqrt{\frac{2\left(H(Z_y)+c^y_l(\phi^s_l)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}\right)\ln(2)+2\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}$$

Finally, since $H(Z_y) = I(X_y;Z_y) + H(Z_y|X_y)$, we have that

$$H(Z_y) + c^y_l(\phi^s_l)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}} = I(X_y;Z_y) + G^y_2.$$

Combining these, we have proven the desired statement of this lemma. □

F.2.6. Combine Lemmas

By combining Lemmas 3, 4, and 5, we have proven the following statement:

Lemma 6. *Let $l \in \{1,\dots,D\}$. If $\phi^s_l \in \Phi$, then for any $\gamma > 0$ and $\delta > 0$, with probability at least $1-\delta$, the following holds:*

$$\mathbb{E}_{X,Y}[\ell((g^s_l \circ \phi^s_l)(X),Y)] - \frac{1}{n}\sum_{i=1}^n \ell((g^s_l \circ \phi^s_l)(x_i),y_i)$$
$$\leq G_3\sqrt{\frac{I(X;Z|Y)\ln(2)+G_2\ln(2)+\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}} + \frac{G_1(\ln|\Phi|)}{\sqrt{n}},$$

where

$$G_1(q) = \frac{\mathcal{L}(f^s)\sqrt{2\gamma|\mathcal{Y}|}}{n^{1/4}}\sqrt{\ln(q)+\ln(2|\mathcal{Y}|/\delta)} + \gamma\mathcal{R}(f^s),$$
$$G_2 = \mathbb{E}_y[c^y_l(\phi^s_l)]\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}} + H(Z|X,Y),$$
$$G_3 = \max_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\sqrt{2|\mathcal{Y}|\mathbb{P}(Z=a^y_k|Y=y)}.$$

Proof. Define the radius of the expected loss R by

$$R = \mathbb{E}_y[R_y] = \mathbb{E}_y\left[\mathbb{E}_{Z,Y}[\ell_l(Z,Y)|Z\notin A_y, Y=y]\right], \tag{31}$$

and the maximum over y of the average training loss per y by

$$\hat{L}(f^s) = \max_{y\in\tilde{\mathcal{Y}}}\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y) = \max_{y\in\tilde{\mathcal{Y}}}\frac{1}{|\tilde{\mathcal{I}}^y|}\sum_{i\in\tilde{\mathcal{I}}^y}\ell(f^s(x_i),y). \tag{32}$$

Let $l \in \{1,\dots,D\}$. By combining Lemmas 3, 4, and 5, if $\phi^s_l \in \Phi$, then for any $\gamma > 0$ and $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$\mathbb{E}_{X,Y}[\ell((g^s_l \circ \phi^s_l)(X),Y)] - \frac{1}{n}\sum_{i=1}^n \ell((g^s_l \circ \phi^s_l)(x_i),y_i)$$
$$\leq \sqrt{2}\sum_{y\in\mathcal{Y}}\left(\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\sqrt{\mathbb{P}(Z=a^y_k,Y=y)}\right)\sqrt{\frac{(I(X_y;Z_y)+G^y_2)\ln(2)+\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}$$

$$+\left(\sum_{y\in\tilde{\mathcal{Y}}}\sqrt{\mathbb{P}(Z\notin\mathcal{Z}^s_{\gamma,y},Y=y)}\frac{\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)}{|\tilde{\mathcal{I}}^y|}\right)\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}+\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)\frac{\gamma R_y}{\sqrt{n}}.$$

Define

$$\tilde{G}_3=\max_{y\in\mathcal{Y}}\sqrt{2}\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\sqrt{\mathbb{P}(Z=a^y_k|Y=y)}.$$

Then, we have

$$\begin{aligned}
&\sqrt{2}\left(\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\sqrt{\mathbb{P}(Z=a^y_k,Y=y)}\right)\\
&=\sqrt{\mathbb{P}(Y=y)}\sqrt{2}\left(\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\sqrt{\mathbb{P}(Z=a^y_k|Y=y)}\right)\\
&\leq\tilde{G}_3\sqrt{\mathbb{P}(Y=y)}
\end{aligned}$$

Using this and Jensen's inequality, we have that

$$\begin{aligned}
&\sqrt{2}\sum_{y\in\mathcal{Y}}\left(\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\sqrt{\mathbb{P}(Z=a^y_k,Y=y)}\right)\sqrt{\frac{(I(X_y;Z_y)+G^y_2)\ln(2)+\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}\\
&\leq\tilde{G}_3\sum_{y\in\mathcal{Y}}\frac{|\mathcal{Y}|}{|\mathcal{Y}|}\sqrt{\mathbb{P}(Y=y)}\sqrt{\frac{(I(X_y;Z_y)+G^y_2)\ln(2)+\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}\\
&\leq\tilde{G}_3|\mathcal{Y}|\sqrt{\sum_{y\in\mathcal{Y}}\frac{1}{|\mathcal{Y}|}\frac{\mathbb{P}(Y=y)\left(I(X_y;Z_y)+G^y_2\right)\ln(2)+\mathbb{P}(Y=y)\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}\\
&=\tilde{G}_3\sqrt{|\mathcal{Y}|}\sqrt{\frac{\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)\left(I(X_y;Z_y)+G^y_2\right)\ln(2)+\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}\\
&=\tilde{G}_3\sqrt{|\mathcal{Y}|}\sqrt{\frac{(I(X;Z|Y)+G_2)\ln(2)+\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}}
\end{aligned}$$

where

$$G_2=\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)G^y_2=\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)\left(c^y_l(\phi^s_l)\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}}+H(Z_y|X_y)\right).$$

Moreover,

$$\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)\frac{\gamma R_y}{\sqrt{n}}=\frac{\gamma}{\sqrt{n}}\sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)R_y=\frac{\gamma R}{\sqrt{n}}.$$

Using Lemma 1 and Jensen's inequality, since $\mathbb{P}_Z(Z\notin\mathcal{Z}^s_{\gamma,y}|Y=y)\leq\frac{\gamma}{\sqrt{n}}$,

$$\begin{aligned}
&\sum_{y\in\tilde{\mathcal{Y}}}\sqrt{\mathbb{P}(Z\notin\mathcal{Z}^s_{\gamma,y},Y=y)}\frac{\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)}{|\tilde{\mathcal{I}}^y|}\\
&=\sum_{y\in\tilde{\mathcal{Y}}}\sqrt{\mathbb{P}(Z\notin\mathcal{Z}^s_{\gamma,y}|Y=y)}\sqrt{\mathbb{P}(Y=y)}\frac{\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)}{|\tilde{\mathcal{I}}^y|}\\
&\leq\hat{L}(f^s)\frac{\sqrt{\gamma}}{n^{1/4}}\sum_{y\in\mathcal{Y}}\frac{|\mathcal{Y}|}{|\mathcal{Y}|}\sqrt{\mathbb{P}(Y=y)}\\
&\leq\hat{L}(f^s)\frac{\sqrt{\gamma}}{n^{1/4}}|\mathcal{Y}|\sqrt{\sum_{y\in\mathcal{Y}}\frac{1}{|\mathcal{Y}|}\mathbb{P}(Y=y)}
\end{aligned}$$

$$= \hat{L}(f^s)\frac{\sqrt{\gamma|\mathcal{Y}|}}{n^{1/4}}$$

Thus, since $R \leq \mathcal{R}(f^s)$ and $\hat{L}(f^s) \leq \mathcal{L}(f^s)$,

$$\frac{G_1(\ln|\Phi|)}{\sqrt{n}} \geq \left(\sum_{y\in\tilde{\mathcal{Y}}} \sqrt{\mathbb{P}(Z \notin \mathcal{Z}^s_{\gamma,y}, Y=y)}\frac{\sum_{i\in\tilde{\mathcal{I}}^y}\ell_l(z_i,y)}{|\tilde{\mathcal{I}}^y|}\right)\sqrt{\frac{2\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}} + \sum_{y\in\mathcal{Y}}\mathbb{P}(Y=y)\frac{\gamma R_y}{\sqrt{n}},$$

where

$$G_1(q) = \frac{\mathcal{L}(f^s)\sqrt{2\gamma|\mathcal{Y}|}}{n^{1/4}}\sqrt{q+\ln(2|\mathcal{Y}|/\delta)} + \gamma\mathcal{R}(f^s).$$

Combining these imply the desired statement. □

F.3. Completing the Proof of Theorem 1 with Key Lemmas

Recall that we have proven the following lemma in the previous subsection:

Lemma 6. *Let $l \in \{1,\dots,D\}$. If $\phi^s_l \in \Phi$, then for any $\gamma > 0$ and $\delta > 0$, with probability at least $1-\delta$, the following holds:*

$$\mathbb{E}_{X,Y}[\ell((g^s_l\circ\phi^s_l)(X),Y)] - \frac{1}{n}\sum_{i=1}^n \ell((g^s_l\circ\phi^s_l)(x_i),y_i) \leq G_3\sqrt{\frac{I(X;Z|Y)\ln(2)+G_2\ln(2)+\ln(2|\Phi||\mathcal{Y}|/\delta)}{n}} + \frac{G_1(\ln|\Phi|)}{\sqrt{n}},$$

where

$$G_1(q) = \frac{\mathcal{L}(f^s)\sqrt{2\gamma|\mathcal{Y}|}}{n^{1/4}}\sqrt{\ln(q)+\ln(2|\mathcal{Y}|/\delta)} + \gamma\mathcal{R}(f^s),$$
$$G_2 = \mathbb{E}_y[c^y_l(\phi^s_l)]\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}} + H(Z|X,Y),$$
$$G_3 = \max_{y\in\mathcal{Y}}\sum_{k=1}^{T_y}\ell_l(a^y_k,y)\sqrt{2|\mathcal{Y}|\mathbb{P}(Z=a^y_k|Y=y)}.$$

Theorem 2 directly follows from Lemma 6; i.e., we complete the proof of Theorem 2 using Lemma 6. Since ϕ^s_l is fixed independently of the training dataset s in Theorem 1, we can invoke Lemma 6 with $\Phi = \{\phi^s_l\}$, with which $|\Phi| = 1$ and $\phi^s_l \in \Phi$. Thus, by noticing that $f^s = g^s_l \circ \phi^s_l$ for any $l \in \{1,\dots,D\}$, Lemma 6 implies the desired statement. □

F.4. Completing the Proof of Theorem 2 with Key Lemmas

We complete the proof of Theorem 2 by extending Lemma 6 in the following.

F.4.1. Finding a likely space of encoder

Fix $l \in \{1,\dots,D\}$ throughout this section. Let $\lambda = \lambda_l$ and $C_\lambda = C_{\lambda,l}$. Recall that $\mathcal{A}_l(s) \in \mathcal{M}_l$ and $|\mathcal{M}_l| < \infty$. For simplicity of notation, we define the random variable A_s by $A_s = \phi^S_l$. For any $q \in \mathcal{M}_l$, we denote

$$p(q) = \mathbb{P}(A_s = q). \tag{33}$$

The entropy of the random variable A_s is given by

$$\mathbb{E}_{A_s}[-\log p(A_s)] = H(A_s).$$

Define the typical subset

$$\Phi_\epsilon^l = \{\phi_l \in \mathcal{M}_l : -\log \mathbb{P}(A_s = \phi_l) - H(A_s) \le \epsilon\}.$$

The following proposition shows that the probability of going outside of the typical subset Φ_ϵ^l is bounded by δ when we take $\epsilon = (1/\lambda)\log(C_\lambda/\delta)$:

Lemma 7. *For any $\lambda > 0$, if we take $\epsilon = (1/\lambda)\ln(C_\lambda/\delta)$, then we have*

$$\mathbb{P}(\phi_l^S \notin \Phi_\epsilon^l) \le \delta, \tag{34}$$

and

$$|\Phi_\epsilon^l| \le 2^{H(\phi_l^S)+\epsilon} = 2^{H(\phi_l^S)+\frac{1}{\lambda}\log\frac{C_\lambda}{\delta}}. \tag{35}$$

Proof. By the definition of the set Φ_ϵ^l, we have

$$\begin{aligned}
\mathbb{P}(A_s \notin \Phi_\epsilon^l) &= \mathbb{P}(q \in \mathcal{M}_l : -\log p(q) \ge H(A_s) + \epsilon) && (36)\\
&= \mathbb{P}(q \in \mathcal{M}_l : -\lambda \log p(q) \ge \lambda H(A_s) + \lambda\epsilon) && (37)\\
&= \mathbb{P}(q \in \mathcal{M}_l : p^{-\lambda}(q) \ge e^{\lambda H(A_s)+\lambda\epsilon}) && (38)\\
&\le e^{-\lambda H(A_s)-\lambda\epsilon} \sum_{q\in\mathcal{M}_l} p^{-\lambda}(q)p(q) && (39)\\
&= \frac{C_\lambda}{e^{\lambda\epsilon}} = \delta. && (40)
\end{aligned}$$

Now we compute the size of Φ_ϵ^l. From the definition of Φ_ϵ^l, we have

$$\begin{aligned}
-\log p(\phi_l) - H(A_s) \le \epsilon &\iff -\log p(\phi_l) \le H(A_s) + \epsilon\\
&\iff -H(A_s) - \epsilon \le \log p(\phi_l)\\
&\iff 2^{-H(A_s)-\epsilon} \le p(\phi_l).
\end{aligned}$$

Using $2^{-H(A_s)-\epsilon} \le p(\phi_l)$,

$$1 \ge \mathbb{P}_s(A_s \in \Phi_\epsilon^l) = \sum_{\phi_l\in\Phi_\epsilon^l} \mathbb{P}(A_s = \phi_l) \ge \sum_{\phi_l\in\Phi_\epsilon^l} 2^{-H(A_s)-\epsilon} = |\Phi_\epsilon^l| 2^{-H(A_s)-\epsilon}.$$

This implies that using $\epsilon = (1/\lambda)\ln(C_\lambda/\delta)$,

$$|\Phi_\epsilon^l| \le 2^{H(A_s)+\frac{1}{\lambda}\ln\frac{C_\lambda}{\delta}}.$$

□

F.4.2. Result with fixed Layer Index

Combining Lemmas 6 and 7 implies the following lemma, which is a main result for a fixed layer index l:

Lemma 8. *Let $l \in \{1, \dots, D\}$. Then, for any $\gamma > 0$ and any $\delta > 0$, with probability at least $1 - \delta$, the following holds:*

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^n \ell(f^s(x_i), y_i) && (41)\\
&\le G_3\sqrt{\frac{\big(I(X;Z|Y) + I(\phi_l^S;S) + G_2 + G_4\big)\ln(2) + \ln(4|\mathcal{Y}|/\delta)}{n}} + \frac{G_1(\tilde{q})}{\sqrt{n}}.
\end{aligned}$$

where $\tilde{q} = (I(\phi_l^S;S) + G_4)\ln(2) + \ln(2)$,

$$G_1(\tilde{q}) = \frac{\mathcal{L}(f^s)\sqrt{2\gamma|\mathcal{Y}|}}{n^{1/4}}\sqrt{\tilde{q} + \ln(2|\mathcal{Y}|/\delta)} + \gamma\mathcal{R}(f^s),$$

$$G_2 = \mathbb{E}_y[c_l^y(\phi_l^s)]\sqrt{\frac{m\ln(\sqrt{n}/\gamma)}{2}} + H(Z|X,Y),$$
$$G_3 = \max_{y\in\mathcal{Y}} \sum_{k=1}^{T_y} \ell_l(a_k^y, y)\sqrt{2|\mathcal{Y}|\mathbb{P}(Z = a_k^y|Y = y)},$$
$$G_4 = \frac{1}{\lambda}\ln\frac{C_\lambda}{\delta} + H(\phi_l^S|S).$$

Proof. Fix $l \in \{1,\dots,D\}$. Let $\lambda > 0$ and $\epsilon = (1/\lambda)\ln(C_\lambda/\delta)$. Using Lemma 6, if $\phi_l^s \in \Phi_\epsilon^l$, then for any $\gamma > 0$ and $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$\mathbb{E}_{X,Y}[\ell((g_l^s \circ \phi_l^s)(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell((g_l^s \circ \phi_l^s)(x_i), y_i) \tag{42}$$
$$\leq G_3\sqrt{\frac{(I(X;Z|Y) + G_2)\ln(2) + \ln(2|\Phi_\epsilon^l||\mathcal{Y}|/\delta)}{n}} + \frac{G_1(\ln|\Phi_\epsilon^l|)}{\sqrt{n}}.$$

From Lemma 7,

$$\mathbb{P}(A_s \notin \Phi_\epsilon^l) \leq \delta$$
$$\mathbb{P}_s(\phi_l^s \notin \Phi_\epsilon^l) \leq \bar{\delta}.$$

Thus, since $\mathbb{P}(A\cap B) \leq \mathbb{P}(B)$ and $\mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(A \mid B)$, we have that

$$\begin{aligned}
&\mathbb{P}_S(\text{Inequality (42) holds})\\
&\geq \mathbb{P}_S(\phi_l^S \in \Phi_\epsilon^l \bigcap \text{Inequality (42) holds})\\
&= \mathbb{P}_S(\phi_l^S \in \Phi_\epsilon^l)\mathbb{P}_S(\text{Inequality (42) holds} \mid \phi_l^S \in \Phi_\epsilon^l)\\
&\geq \mathbb{P}_S(\phi_l^S \in \Phi_\epsilon^l)(1-\delta)\\
&\geq (1-\delta)(1-\delta) = 1 - 2\delta + \delta^2 \geq 1 - 2\delta.
\end{aligned}$$

Therefore, by setting $\delta = \frac{\delta'}{2}$, we have that for any $\delta' > 0$,

$$\mathbb{P}_S(\text{Eq (42) holds}) \geq 1 - \delta'.$$

In other words, for any $\gamma > 0$ and $\delta > 0$, with probability at least $1-\delta$, the following holds:

$$\mathbb{E}_{X,Y}[\ell((g_l^s \circ \phi_l^s)(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell((g_l^s \circ \phi_l^s)(x_i), y_i) \tag{43}$$
$$\leq G_3\sqrt{\frac{(I(X;Z|Y) + G_2)\ln(2) + \ln(4|\Phi_\epsilon^l||\mathcal{Y}|/\delta)}{n}} + \frac{G_1(\ln 2|\Phi_\epsilon^l|)}{\sqrt{n}}.$$

From Lemma 7, we have $|\Phi_\epsilon^l| \leq 2^{H(\phi_l^S) + \frac{1}{\lambda}\ln\frac{C_\lambda}{\delta}}$ and thus

$$\ln(4|\Phi_\epsilon^l||\mathcal{Y}|/\delta) = \ln(|\Phi_\epsilon^l|) + \ln(4|\mathcal{Y}|/\delta) \leq \left(H(\phi_l^S) + \frac{1}{\lambda}\ln\frac{C_\lambda}{\delta}\right)\ln(2) + \ln(4|\mathcal{Y}|/\delta).$$

From the definition of the entropy, conditional entropy, and mutual information, we have that

$$H(\phi_l^S) = I(\phi_l^S; S) + H(\phi_l^S|S).$$

Using this,

$$H(\phi_l^S) + \frac{1}{\lambda}\ln\frac{C_\lambda}{\delta} = I(\phi_l^S; S) + G_4.$$

By combining these and noticing that $f^s = g_l^s \circ \phi_l^s$ for any $l \in \{1, \dots, D\}$, we have that for any $\gamma > 0$ and $\delta > 0$, with probability at least $1 - \delta$, the following holds:

$$\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i) \tag{44}$$
$$\leq G_3 \sqrt{\frac{\big(I(X; Z|Y) + I(\phi_l^S; S) + G_2 + G_4\big)\ln(2) + \ln(4|\mathcal{Y}|/\delta)}{n}} + \frac{G_1(\tilde{q})}{\sqrt{n}}.$$

□

F.4.3. Completing the Proof

We complete the proof of Theorem 2 using Lemma 8. Let $\gamma_l > 0$ and $\lambda_l > 0$ for all $l \in \{1, 2, \dots, D+1\}$. Recall that $f^s = g_l^s \circ \phi_l^s$ for any $l \in \{1, \dots, D\}$. Thus, by making the dependence of the layer index l explicit, Lemma 8 states that for any $\delta > 0$ and (fixed) $l \in \{1, \dots, D\}$, with probability at least $1 - \delta$,

$$\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i) \tag{45}$$
$$\leq G_3^l \sqrt{\frac{\big(I(X; Z_l^s|Y) + I(\phi_l^S; S) + G_2^l + G_4^l\big)\ln(2) + \ln(4|\mathcal{Y}|/\delta)}{n}} + \frac{G_1^l(\tilde{q})}{\sqrt{n}},$$

where $\tilde{q} = (I(\phi_l^S; S) + G_4)\ln(2) + \ln(2)$,

$$G_1^l(q) = \frac{\mathcal{L}(f^s)\sqrt{2\gamma_l|\mathcal{Y}|}}{n^{1/4}}\sqrt{q + \ln(2|\mathcal{Y}|/\delta)} + \gamma_l \mathcal{R}(f^s),$$
$$G_2^l = \mathbb{E}_y[c_l^y(\phi_l^s)]\sqrt{\frac{m\ln(\sqrt{n}/\gamma_l)}{2}} + H(Z_l^s|X, Y).$$
$$G_3^l = \max_{y \in \mathcal{Y}} \sum_{k=1}^{T_y} \ell_l(a_k^y, y)\sqrt{2|\mathcal{Y}|\mathbb{P}(Z = a_k^y|Y = y)},$$
$$\tilde{G}_4^l = \frac{1}{\lambda_l}\ln\frac{C_{\lambda_l, l}}{\delta} + H(\phi_l^S|S).$$

We now consider the case of $l = D+1$. Let $l = D+1$ and $\lambda_{D+1} > 0$. Fix $f = \phi_{D+1} \in \Phi_\epsilon^{D+1}$ with $\epsilon = (1/\lambda)\ln(C_{\lambda_{D+1}, D+1}/\delta)$. Then, by using Hoeffding's inequality, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathbb{E}_{X,Y}[\ell(f(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) \leq \mathcal{R}(f)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

By taking union bounds over elements of Φ_ϵ^{D+1}, this implies that for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f \in \Phi_\epsilon^{D+1}$,

$$\mathbb{E}_{X,Y}[\ell(f(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) \leq \mathcal{R}(f)\sqrt{\frac{\ln(|\Phi_\epsilon^{D+1}|/\delta)}{2n}}.$$

This implies that for any $\delta > 0$, if $\phi_{D+1}^s \in \Phi_\epsilon^{D+1}$, then with probability at least $1 - \delta$,

$$\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i) \leq \mathcal{R}(f^s)\sqrt{\frac{\ln(|\Phi_\epsilon^{D+1}|/\delta)}{2n}}. \tag{46}$$

Here, from Lemma 7, we have that

$$\mathbb{P}(\phi_{D+1}^S \notin \Phi_\epsilon^{D+1}) \leq \delta.$$

Since $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$ and $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(A \mid B)$, we have that

$$\begin{aligned}
&\mathbb{P}_S(\text{Inequality (46) holds}) \\
&\geq \mathbb{P}_S(\phi_{D+1}^S \in \Phi_\epsilon^{D+1} \bigcap \text{Inequality (46) holds}) \\
&= \mathbb{P}_S(\phi_{D+1}^S \in \Phi_\epsilon^{D+1})\mathbb{P}_S(\text{Inequality (46) holds} \mid \phi_{D+1}^S \in \Phi_\epsilon^{D+1}) \\
&\geq \mathbb{P}_S(\phi_{D+1}^S \in \Phi_\epsilon^{D+1})(1-\delta) \\
&\geq (1-\delta)(1-\delta) \\
&\geq 1-2\delta.
\end{aligned}$$

Therefore, by setting $\delta = \frac{\delta'}{2}$, we have that for any $\delta' > 0$,

$$\mathbb{P}_S(\text{Eq (46) holds}) \geq 1-\delta'.$$

In other words, for any $\delta' > 0$, with probability at least $1-\delta'$,

$$\mathbb{E}_{X,Y}[\ell(f^s(X),Y)] - \frac{1}{n}\sum_{i=1}^n \ell(f^s(x_i),y_i) \leq \mathcal{R}(f^s)\sqrt{\frac{\ln(2|\Phi_\epsilon^{D+1}|/\delta')}{2n}}. \tag{47}$$

Here, from Lemma 7, we have that

$$|\Phi_\epsilon^{D+1}| \leq 2^{H(\phi_{D+1}^S)+\frac{1}{\lambda_{D+1}}\log\frac{C_{\lambda_{D+1},D+1}}{\delta}}.$$

Substituting this,

$$\begin{aligned}
\ln(2|\Phi_\epsilon^{D+1}|/\delta') &= \ln(|\Phi_\epsilon^{D+1}|) + \ln(2/\delta') \\
&\leq \left(H(\phi_{D+1}^S) + \frac{1}{\lambda_{D+1}}\log\frac{C_{\lambda_{D+1},D+1}}{\delta}\right)\ln(2) + \ln(2/\delta')
\end{aligned}$$

Using $H(\phi_{D+1}^S) = I(\phi_{D+1}^S;S) + H(\phi_{D+1}^S|S)$,

$$H(\phi_{D+1}^S) + \frac{1}{\lambda_{D+1}}\log\frac{C_{\lambda_{D+1},D+1}}{\delta} = I(\phi_{D+1}^S;S) + \tilde{G}_4^{D+1}.$$

Substituting these into (47), we have that for any $\delta > 0$, with probability at least $1-\delta$,

$$\mathbb{E}_{X,Y}[\ell(f^s(X),Y)] - \frac{1}{n}\sum_{i=1}^n \ell(f^s(x_i),y_i) \leq \mathcal{R}(f^s)\sqrt{\frac{\left(I(\phi_{D+1}^S;S) + \tilde{G}_4^{D+1}\right)\ln(2) + \ln(2/\delta)}{2n}}. \tag{48}$$

By combining (45) and (48) with union bounds over $\mathcal{D}$, we have that for any $\delta > 0$ and $\mathcal{D} \subseteq \{1,2,\ldots,D+1\}$, with probability at least $1-\delta$, the following holds for all $l \in \mathcal{D}$:

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\ell(f^s(X),Y)] - \frac{1}{n}\sum_{i=1}^n \ell(f^s(x_i),y_i) \\
&\leq \mathbb{1}\{l \neq D_+\}\left(G_3^l\sqrt{\frac{\left(I(X;Z_l^s|Y) + I(\phi_l^S;S) + G_2^l + G_4^l\right)\ln(2) + \ln(4|\mathcal{Y}||\mathcal{D}|/\delta)}{n}} + \frac{G_1^l}{\sqrt{n}}\right) \\
&\quad + \mathbb{1}\{l = D_+\}\mathcal{R}(f^s)\sqrt{\frac{\left(I(\phi_{D+1}^S;S) + G_4^{D+1}\right)\ln(2) + \ln(2/\delta)}{2n}},
\end{aligned} \tag{49}$$

where $D_+ = D+1$,

$$G_4^l = \frac{1}{\lambda_l}\ln\frac{C_{\lambda_l,l}|\mathcal{D}|}{\delta} + H(\phi_l^S|S).$$

Since the right-hand side of this inequality holds for all $l \in \mathcal{D}$ and the left-hand side does not depend on l, this implies that for any $\delta > 0$ and $\mathcal{D} \subseteq \{1, 2, \dots, D+1\}$, with probability at least $1-\delta$,

$$\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i) \leq \min_{l \in \mathcal{D}} Q_l, \tag{50}$$

$$\text{where } Q_l = \begin{cases} G_3^l \sqrt{\frac{\left(I(X;Z_l^s|Y) + I(\phi_l^S;S)\right)\ln(2) + \widehat{\mathcal{G}}_2^l}{n}} + \frac{G_1^l(\zeta)}{\sqrt{n}} & \text{if } l \leq D \\ \mathcal{R}(f^s)\sqrt{\frac{I(\phi_l^S;S)\ln(2) + \check{\mathcal{G}}_2^l}{2n}} & \text{if } l = D+1, \end{cases}$$

where $\zeta = (I(\phi_l^S;S) + G_4^l)\ln(2) + \ln(2|\mathcal{D}|)$, $\widehat{\mathcal{G}}_2^l = \left(G_2^l + G_4^l\right)\ln(2) + \ln(4|\mathcal{Y}||\mathcal{D}|/\delta)$, $\check{\mathcal{G}}_2^l = G_4^l \ln(2) + \ln(2/\delta)$, and $G_4^l = \frac{1}{\lambda_l}\ln\frac{C_{\lambda_l,l}|\mathcal{D}|}{\delta} + H(\phi_l^S|S)$.

□

F.5. Proof of Remark 1

The desired statement follows from

$$I(\theta_l^S;S) + H(\theta_l^S|S) = H(\theta_l^S) \geq H(\phi_{l,\theta_l^S}) = H(\phi_l^S) = I(\phi_l^S;S) + H(\phi_l^S|S),$$

where the inequality holds because all the randomness of ϕ_{l,θ_l^S} comes from the randomness of $\theta_l^S = \mathcal{A}_l^\theta \circ S$ (where $\mathcal{A}_l^\theta$ is the version of $\mathcal{A}_l$ that outputs the parameter vector instead of the encoder function), and because one ϕ_{l,θ_l^S} corresponds to one or more θ_l^S; i.e., we have $\phi_{l,\theta_l^s} = \phi_{l,\bar{\theta}_l^s}$ whenever $\theta_l^s = \bar{\theta}_l^s$ and it is possible that $\phi_{l,\theta_l^s} = \phi_{l,\bar{\theta}_l^s}$ for $\theta_l^s \neq \bar{\theta}_l^s$. In other words, the desired statement does not hold only if $\phi_{l,\theta_l^s} \neq \phi_{l,\bar{\theta}_l^s}$ for some $\theta_l^s = \bar{\theta}_l^s$, which is not the case.

□

F.6. Proof of Corollary 1

Proof. Set $\phi_l^s = \mathcal{E}_l[\tilde{\phi}_l^s] \circ \tilde{\phi}_l^s$. Then, Theorems 1–2 hold true for this choice of encoder ϕ_l^s since this does not violate any assumption of Theorems 1–2. Thus, Theorems 1–2 hold with eq. (6) and eq. (7) in their original forms: i.e.,

$$\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i) \leq \hat{Q}_l, \text{ and,} \tag{51}$$

$$\mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i) \leq \min_{l \in \mathcal{D}} Q_l,$$

Since $\mathbb{P}(|\ell((g_l^s \circ \tilde{\phi}_l^s)(X), Y) - \ell((g_l^s \circ \mathcal{E}_l[\tilde{\phi}_l^s] \circ \tilde{\phi}_l^s)(X), Y)| \leq C_l) = \mathbb{P}(|\ell(\tilde{f}^s(X), Y) - \ell(f^s(X), Y)| \leq C_l) = 1$, we have that with probability one,

$$\ell(\tilde{f}^s(x_i), y_i) = \ell(f^s(x_i), y_i) + (\ell(\tilde{f}^s(x), y) - \ell(f^s(x_i), y_i)) \leq \ell(f^s(x_i), y_i) + C_l. \tag{52}$$

Thus, with probability one,

$$\mathbb{E}_{X,Y}[\ell(\tilde{f}^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(\tilde{f}^s(x_i), y_i) \leq \mathbb{E}_{X,Y}[\ell(f^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f^s(x_i), y_i) + 2C_l. \tag{53}$$

Combining eq. (51) and eq. (53) with union bounds concludes that Theorems 1–2 hold when we replace eq. (6) and eq. (7) by

$$\mathbb{E}_{X,Y}[\ell(\tilde{f}^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(\tilde{f}^s(x_i), y_i) \leq \hat{Q}_l + 2C_l, \text{ and,} \tag{54}$$

$$\mathbb{E}_{X,Y}[\ell(\tilde{f}^s(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(\tilde{f}^s(x_i), y_i) \leq \min_{l \in \mathcal{D}} Q_l + 2C_l,$$

Finally, the values of $\hat{Q}_l$ and Q_l are finite since $|\mathcal{Z}_l^s| < \infty$ and $|\mathcal{M}_l| < \infty$; e.g., $|\mathcal{Z}_l^s| < \infty$ implies that $I(X;Z_l^s|Y) < \infty$. Thus, if $C_{\mathcal{E}} < \infty$, we have $\hat{Q}_l + 2C_l < \infty$ and $\min_{l \in \mathcal{D}} Q_l + 2C_l < \infty$. □

F.7. Proof of Proposition 1

Proof. Let l be fixed and $\phi^s = \phi^s_l$. For deterministic neural networks, the intermediate output Z is a deterministic function of the input X, i.e. $Z = \phi^s(X)$. In this case the conditional mutual information between X and Z simplifies to the conditional entropy of Z:

$$I(X, Z|Y) = H(Z|Y) = H(\phi^s(X)|Y). \tag{55}$$

It has been proven in (Amjad & Geiger, 2019) that if X has absolutely continuous component, which has continous density on a compact set, and the activation is bi-Lipschitz or continuous differentiable with strictly positive derivative, then the entropy of $\phi^s(X)$ is infinite.

In the following, we first give some simple examples with ReLU activation where the entropy of $\phi^s(X)$ is finite for an initial intuition, and then we generalize the examples for more practical settings. Finally, we discuss generality and practicality of our construction.

Consider an arbitrary (continuous or discrete) distribution such that the distribution of $X|Y$ consists of several components (which may correspond to further subclasses). For simplicity, we assume that there are two components $\mathcal{C}_1$ and $\mathcal{C}_2$. We start with the linearly separable case where $\mathcal{C}_1$ and $\mathcal{C}_2$ are separated by a hyperplane $ax + b = 0$ with margin at least r. In other words $ax + b \geq r$ for $x \in \mathcal{C}_1$, and $ax + b \leq -r$ for $x \in \mathcal{C}_2$. Then the following simple one layer network

$$\sigma(ax + b + c) - \sigma(ax + b - c), \quad 0 < c \leq r \tag{56}$$

maps $x \in \mathcal{C}_1$ to $2c$ and $\mathcal{C}_2$ to 0. Thus, the output $Z = \phi^s(X)$ follows a Bernoulli distribution, which has bounded entropy. Thus, we have $I(X, Z|Y) < \infty$.

More generally, if $\mathcal{C}_1$ and $\mathcal{C}_2$ are separable, with margin at least r using some metric $d(\mathcal{C}_1, \mathcal{C}_2) \geq r$, we can take a Lipschitz function g w.r.t. this metric d with lipschitz constant $1/r$ such that it equals 0 on C_1 and equals 1 on C_2. By the universal approximation theory, ReLU neural network can approximate arbitrary continuous function to arbitrary precision as we increase the network size. In particular, there exists a finite-size ReLU neural network N such that $|N(x) - g(x)| \leq 1/8$. As a consequence, we have $N(x) \leq 1/8$ for $x \in \mathcal{C}_1$ and $N(x) \geq 7/8$ for $x \in \mathcal{C}_2$. We consider the following neural network:

$$\sigma(N - 1/2 + c) - \sigma(N - 1/2 - c), \quad 0 < c < 3/8, \tag{57}$$

which maps $x \in \mathcal{C}_1$ to $2c$ and $\mathcal{C}_2$ to 0. Thus, the output $Z = \phi^s(X)$ follows a Bernoulli distribution, which has bounded entropy. Thus, we have $I(X, Z|Y) < \infty$. Since the distribution in this general example is arbitrary except for the separable components, there exists infinitely many such distributions.

Finally, we observe that these examples are general and practical. First, the above proof works for any finite number of separable components instead of two components. Second, it is also observed in practice that trained neural networks behave like these examples discussed above, which maps different class to different points; this is sometimes referred as a neural collapse phenomenon (Papyan et al., 2020). □

F.8. Proof of Proposition 2

We will use the following lemma to prove Proposition 2:

Lemma 9. *Let $v_1, \ldots, v_T \in \mathbb{R}$ such that $0 \leq v_k \leq Ce^{-(k/\beta)^\alpha}$ for some constants $\alpha \geq 1$ and $\beta, C > 0$. Then,*

$$\sum_{k=1}^{T} \sqrt{v_k} \leq \sum_{k=1}^{\lceil \tilde{\beta} \rceil} \sqrt{v_k} + \frac{C\tilde{\beta}}{\alpha e}$$

where $\tilde{\beta} = 2^{1/\alpha}\beta$.

Proof. Using the condition on v_k,

$$\sqrt{v_k} \leq \sqrt{Ce^{-(k/\beta)^\alpha}} = \sqrt{C}\sqrt{e^{-(k/\beta)^\alpha}} = \sqrt{C}e^{-\frac{k^\alpha}{2\beta^\alpha}} = \sqrt{C}e^{-(k/\tilde{\beta})^\alpha}$$

Then,

$$\sum_{k=1}^{T}\sqrt{v_k}=\sum_{k=1}^{\lceil\tilde{\beta}\rceil}\sqrt{v_k}+\sum_{k=\lceil\tilde{\beta}\rceil+1}^{T}\sqrt{v_k}\leq\sum_{k=1}^{\lceil\tilde{\beta}\rceil}\sqrt{v_k}+\sqrt{C}\sum_{k=\lceil\tilde{\beta}\rceil+1}^{T}e^{-(k/\tilde{\beta})^{\alpha}}$$

We now bound the last term by using integral as

$$\sum_{k=\lceil\tilde{\beta}\rceil+1}^{T}e^{-(k/\tilde{\beta})^{\alpha}}\leq\int_{\tilde{\beta}}^{\infty}e^{-(q/\tilde{\beta})^{\alpha}}dq=\frac{\tilde{\beta}}{\alpha}\int_{(\tilde{\beta}/\tilde{\beta})^{\alpha}}^{\infty}t^{\frac{1}{\alpha}-1}e^{-t}dt=\frac{\tilde{\beta}}{\alpha}\int_{1}^{\infty}t^{\frac{1}{\alpha}-1}e^{-t}dt$$

Here, since $\alpha\geq 1$ and $t\geq 1$ in the integral, we have $t^{\frac{1}{\alpha}-1}\leq 1$ in the integral. Thus,

$$\int_{1}^{\infty}t^{\frac{1}{\alpha}-1}e^{-t}dt\leq\int_{1}^{\infty}e^{-t}dt=e^{-1}.$$

By combining these, we have

$$\sum_{k=1}^{T}\sqrt{v_k}\leq\sum_{k=1}^{\lceil\tilde{\beta}\rceil}\sqrt{v_k}+\frac{C\tilde{\beta}}{\alpha e}$$

□

Using Lemma 9, we complete the proof of Proposition 2 in the following:

Proof of Proposition 2. Let $y\in\mathcal{Y}$ and $l\in\{1,\ldots,D\}$. To invoke Lemma 9, we rearrange the expression of G_3^l as

$$\begin{aligned}G_3^l&=\max_{y\in\mathcal{Y}}\sum_{k=1}^{T_y^l}\ell(g_l^s(a_k^{l,y}),y)\sqrt{2|\mathcal{Y}|\mathbb{P}(Z_{l,y}^s=a_k^{l,y})}\\&\leq\sqrt{2|\mathcal{Y}|}\max_{y\in\mathcal{Y}}\sum_{k=1}^{T_y^l}\sqrt{\ell(g_l^s(a_k^{l,y}),y)^2\mathbb{P}(Z_{l,y}^s=a_k^{l,y})}\end{aligned}$$

Then, we invoke Lemma 9 with $v_k=v_k^{(y)}=\ell(g_l^s(a_k^{l,y}),y)^2\mathbb{P}(Z_{l,y}^s=a_k^{l,y})$, where we define $v_k^{(y)}=v_k(y)$. This implies that

$$\sum_{k=1}^{T_y^l}\sqrt{\ell(g_l^s(a_k^{l,y}),y)^2\mathbb{P}(Z_{l,y}^s=a_k^{l,y})}\leq\sum_{k=1}^{\lceil\tilde{\beta}_y\rceil}\sqrt{v_k^{(y)}}+\frac{C_y\tilde{\beta}_y}{\alpha_y e},$$

where $\tilde{\beta}_y=2^{1/\alpha_y}\beta_y$. Thus,

$$G_3^l\leq\sqrt{2|\mathcal{Y}|}\max_{y\in\mathcal{Y}}\left(\sum_{k=1}^{\lceil\tilde{\beta}_y\rceil}\sqrt{v_k^{(y)}}+\frac{C_y\tilde{\beta}_y}{\alpha_y e}\right)$$

□

F.9. Proof of Proposition 3

Proof. Let $l\in\{1,2,\ldots,D+1\}$ and let us write $\lambda=\lambda_l$ and $C_\lambda=C_{\lambda_l,l}$. We first note that the value of C_λ is always bounded as

$$C_\lambda\leq\sum_{q\in\mathcal{M}_l}(\mathbb{P}(\phi_l^S=q))^{1-\lambda}\leq|\mathcal{M}_l|\left(\frac{1}{|\mathcal{M}_l|}\sum_{q\in\mathcal{M}_l}\mathbb{P}(\phi_l^S=q)\right)^{1-\lambda}=|\mathcal{M}_l|^{\lambda},\tag{58}$$

which is a very loose bound and we will provide tighter bounds in below. Before proceeding to the proof, we recall the following bounds. For $a > 1$ we have

$$\frac{1}{a-1} \leq \int_1^\infty \frac{dx}{x^a} \leq \sum_{i=1}^\infty \frac{1}{i^a} \leq 1 + \int_1^\infty \frac{dx}{x^a} = \frac{a}{a-1}, \tag{59}$$

$$\sum_{i=1}^\infty \frac{\ln(i)}{i^a} \leq \frac{\ln(2)}{2^a} + \frac{\ln(3)}{3^a} + \int_3^\infty \frac{\ln(x)dx}{x^a} = \frac{\ln(2)}{2^a} + \frac{\ln(3)}{3^a} + \frac{3^{1-a}((a-1)\ln(3)+1)}{(a-1)^2}, \tag{60}$$

$$\sum_{i=1}^\infty \frac{\ln(i)}{i^a} \geq \frac{\ln(2)}{2^a} + \int_3^\infty \frac{\ln(x)dx}{x^a} \geq \frac{\ln(2)}{2^a} + \frac{3^{1-a}((a-1)\ln(3)+1)}{(a-1)^2}. \tag{61}$$

For $a < 1$, we have

$$\frac{N^{1-a}-1}{1-a} = \int_1^N \frac{dx}{x^a} \leq \sum_{i=1}^N \frac{1}{i^a} \leq 1 + \int_1^N \frac{dx}{x^a} = \frac{N^{1-a}-a}{1-a} \leq \frac{N^{1-a}}{1-a}. \tag{62}$$

In the first case, we have

$$C_\lambda \leq \sum_{i=1}^N p_i^{1-\lambda} \leq \sum_{i=1}^N \frac{C^{1-\lambda}}{i^{\alpha(1-\lambda)}} \leq C^{1-\lambda} \frac{\alpha(1-\lambda)}{\alpha(1-\lambda)-1}, \tag{63}$$

which is bounded and independent of N. For the entropy, we notice that on $[0,1]$ the function $-p\ln p$ is non-negative, increasing on $[0, 1/e]$ and decreasing on $[1/e, 1]$.

$$H(A_s) = \sum_{i=1}^N -p_i \ln(p_i) \leq \sum_{p_i > 1/e} \frac{1}{e} + \sum_{p_i < 1/e} \frac{C}{i^\alpha} \ln \frac{i^\alpha}{C} \tag{64}$$

$$\leq 1 + \sum_{i \geq 1} \frac{C\alpha \ln i}{i^\alpha} \leq 1 + C\alpha \left(\frac{\ln(2)}{2^\alpha} + \frac{\ln(3)}{3^\alpha} + \frac{3^{1-\alpha}((\alpha-1)\ln(3)+1)}{(\alpha-1)^2} \right). \tag{65}$$

In the second case, the normalization constant Z diverges with N,

$$Z = \sum_{i=1}^N \frac{c_i}{i^\alpha} \leq \sum_{i=1}^N \frac{C}{i^\alpha} \leq C \left(1 + \int_1^N \frac{dx}{x^\alpha} \right) \leq C \left(\frac{N^{1-\alpha}}{1-\alpha} \right). \tag{66}$$

And using $c_i \geq c$, we have a lower bound for Z

$$Z \geq c \left(\int_1^N \frac{dx}{x^\alpha} \right) \geq \frac{c(N^{1-\alpha}-1)}{1-\alpha}. \tag{67}$$

Thus Z is of order $N^{1-\alpha}$, i.e. $Z = \Omega(N^{1-\alpha})$.

We recall the formula of C_λ from (21)

$$\ln C_\lambda = \ln \left(\sum_{i=1}^N p_i^{1-\lambda} \right) - \lambda H(A_s). \tag{68}$$

For the first term on the righthand side of (68), we have

$$\begin{aligned} \ln \left(\sum_{i=1}^N p_i^{1-\lambda} \right) &= \ln \left(\sum_{i=1}^N \left(\frac{c_i}{Z i^\alpha} \right)^{1-\lambda} \right) = -(1-\lambda)\ln(Z) + \ln \left(\sum_{i=1}^N \frac{c_i^{1-\lambda}}{i^{(1-\lambda)\alpha}} \right) \\ &= -(1-\lambda)\ln(N^{1-\alpha}) + \ln(N^{1-(1-\lambda)\alpha}) + \mathcal{E}_0 = \lambda \ln N + \mathcal{E}_0, \end{aligned} \tag{69}$$

where

$$|\mathcal{E}_0 - (\ln(1-(1-\lambda)\alpha) - (1-\lambda)\ln(1-\alpha))| \leq (1-\lambda)\ln(C/c) \tag{70}$$

Next we compute the entropy $H(A_s)$ and show it diverges as $\ln N$.

$$\begin{aligned} H(A_s) &= -\sum_{i=1}^{N} p_i \ln(p_i) = -\sum_{i=1}^{N} p_i \ln \frac{c_i}{i^\alpha Z} = \alpha \sum_{i=1}^{N} p_i \ln(i) + \ln(Z) - \sum_{i=1}^{N} p_i \ln(c_i) \\ &= \alpha \sum_{i=1}^{N} p_i \ln(i) + (1-\alpha)\ln N + \mathcal{E}_1, \end{aligned} \tag{71}$$

where

$$\ln(c/C) - \ln(1-\alpha) \leq \mathcal{E}_1 \leq \ln(C/c) - \ln(1-\alpha) \tag{72}$$

To compute the first term on the righthand side of (71), we introduce

$$s_i = \frac{c_1}{1^\alpha} + \frac{c_2}{2^\alpha} + \cdots + \frac{c_i}{i^\alpha}, \quad 1 \leq i \leq N, \tag{73}$$

then $s_N = Z$. Next we can do a summation by part

$$\begin{aligned} \sum_{i=1}^{N} p_i \ln(i) &= \frac{1}{Z}\sum_{i=1}^{N} \frac{c_i}{i^\alpha}\ln(i) = \frac{1}{Z}\sum_{i=1}^{N}(s_i - s_{i-1})\ln(i) \\ &= \frac{1}{Z}\sum_{i=1}^{N-1} s_i(\ln(i) - \ln(i+1)) + \frac{s_N \ln N}{Z} \\ &= \frac{1}{Z}\sum_{i=1}^{N-1} s_i(\ln(i) - \ln(i+1)) + \ln N \end{aligned} \tag{74}$$

The same as in (66), we have $|s_i| \leq Ci^{1-\alpha}/(1-\alpha)$. Moreover $\ln(1+1/i) \leq 1/i$. Thus the first term on the righthand side of (74) can be bounded as

$$\left|\frac{1}{Z}\sum_{i=1}^{N-1} s_i(\ln(i) - \ln(i+1))\right| \leq \frac{1}{Z}\sum_{i=1}^{N} C\left(\frac{i^{1-\alpha}}{1-\alpha}\right)\frac{1}{i} \leq \frac{C}{c(1-\alpha)} \tag{75}$$

By plugging (74) and (75) into (71), we conclude the following bound on the entropy

$$H(A_s) = \ln(N) + \mathcal{E}_2. \tag{76}$$

where

$$\ln(c/C) - \ln(1-\alpha) - \frac{C}{c(1-\alpha)} \leq \mathcal{E}_2 \leq \ln(C/c) - \ln(1-\alpha) + \frac{C}{c(1-\alpha)} \tag{77}$$

The two estimates (69) and (76) together imply that $C_\lambda = \mathcal{E}_0 - \lambda\mathcal{E}_2$, and

$$|C_\lambda - (\ln(1-(1-\lambda)\alpha) - (1-2\lambda)\ln(1-\alpha))| \leq (2-\lambda)\ln(C/c) + \frac{C}{c(1-\alpha)} \tag{78}$$

□