entropy

MDPI

Article

Towards a Unified Theory of Learning and Information

Ibrahim Alabdulmohsin

Google Research, 8002 Zürich, Switzerland; ibomohsin@google.com

Received: 10 February 2020; Accepted: 6 April 2020; Published: 13 April 2020

Abstract: In this paper, we introduce the notion of "learning capacity" for algorithms that learn from data, which is analogous to the Shannon channel capacity for communication systems. We show how "learning capacity" bridges the gap between statistical learning theory and information theory, and we will use it to derive generalization bounds for finite hypothesis spaces, differential privacy, and countable domains, among others. Moreover, we prove that under the Axiom of Choice, the existence of an empirical risk minimization (ERM) rule that has a vanishing learning capacity is equivalent to the assertion that the hypothesis space has a finite Vapnik–Chervonenkis (VC) dimension, thus establishing an equivalence relation between two of the most fundamental concepts in statistical learning theory and information theory. In addition, we show how the learning capacity of an algorithm provides important qualitative results, such as on the relation between generalization and algorithmic stability, information leakage, and data processing. Finally, we conclude by listing some open problems and suggesting future directions of research.

Keywords: statistical learning theory; information theory; entropy; parameter estimation; learning systems; privacy; prediction methods

1. Introduction

1.1. Generalization Risk

A central goal when learning from data is to strike a balance between underfitting and overfitting. Mathematically, this requirement can be translated into an optimization problem with two competing objectives. First, we would like the learning algorithm to produce a hypothesis (i.e., an answer) that performs well on the empirical sample. This goal can be easily achieved by using a *rich* hypothesis space that can "explain" any observations. Second, we would like to guarantee that the performance of the hypothesis on the empirical data (a.k.a. training error) is a good approximation of its performance with respect to the unknown underlying distribution (a.k.a. test error). This goal can be achieved by *limiting* the complexity of the hypothesis space. The first condition mitigates underfitting while the latter condition mitigates overfitting.

Formally, suppose we have a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ that receives a sample $\mathbf{s} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$, which comprises of m i.i.d. observations $\mathbf{z}_i \sim p(z)$, and uses $\mathbf{s}$ to select a hypothesis $\mathbf{h} \in \mathcal{H}$. Let l be a loss function defined on the product space $\mathcal{Z} \times \mathcal{H}$. For instance, l can be the mean-square-error (MSE) in regression or the 0–1 error in classification. Then, the goal of learning from data is to select a hypothesis $\mathbf{h} \in \mathcal{H}$ such that its *true risk* $R(\mathbf{h})$, defined by

$$R(h) = \mathbb{E}_{\mathbf{z} \sim p(z)}[l(\mathbf{z}, h)], \tag{1}$$

is small. However, this optimization problem is often difficult to solve exactly since the underlying distribution of observations $p(z)$ is seldom known. Rather, because the true risk $R(\mathbf{h})$ can be decomposed into a sum of two terms:

$$R(\mathbf{h}) = \big[R_{\mathbf{s}}(\mathbf{h})\big] + \big[R(\mathbf{h}) - R_{\mathbf{s}}(\mathbf{h})\big],$$

where $R_{\mathbf{s}}(\mathbf{h}) = \mathbb{E}_{\mathbf{z}\sim\mathbf{s}}[l(\mathbf{z},\mathbf{h})] \doteq (1/m)\sum_{z\in\mathbf{s}} l(z,\mathbf{h})$, both terms can be tackled separately. The first term in the equation above corresponds to the *empirical risk* on the training sample $\mathbf{s}$. The second term corresponds to the *generalization risk*. Hence, by minimizing both terms, one obtains a learning algorithm whose true risk is small.

Minimizing the empirical risk can be achieved using tractable approximations to the *empirical risk minimization* (ERM) procedure, such as stochastic convex optimization [1,2]. However, the generalization risk is often difficult to deal with directly because the underlying distribution is often unknown. Instead, it is a common practice to bound it *analytically*. By establishing analytical conditions for generalization, one hopes to design better learning algorithms that both perform well empirically and generalize as well into the future.

Several methods have been proposed in the past for bounding the generalization risk of learning algorithms. Some examples of popular approaches include uniform convergence, algorithmic stability, Rademacher and Gaussian complexities, and the PAC–Bayesian framework [3–7].

The proliferation of such bounds can be understood upon noting that the generalization risk of a learning algorithm is influenced by multiple factors, such as the domain $\mathcal{Z}$, the hypothesis space $\mathcal{H}$, and the mapping from $\mathcal{Z}$ to $\mathcal{H}$. Hence, one may derive new generalization bounds by imposing conditions on any of such components. For example, the Vapnik–Chervonenkis (VC) theory derives generalization bounds by assuming constraints on $\mathcal{H}$ whereas stability bounds, e.g., [6,8,9], are derived by assuming constraints on the mapping from $\mathcal{Z}$ to $\mathcal{H}$.

Rather than showing that certain conditions are sufficient for generalization, we will establish in this paper conditions that are both *necessary and sufficient*. More precisely, we will show that the " uniform" generalization risk of a learning algorithm is an *information-theoretic* characterization. In particular, it is *equal* to the total variation distance between the joint distribution of the hypothesis $\mathbf{h}$ and a single random training example $\hat{\mathbf{z}} \sim \mathbf{s}$, on one hand, and the product of their marginal distributions, on the other hand. Hence, it is analogous to the mutual information between $\mathbf{h}$ and $\hat{\mathbf{z}}$. Since uniform generalization is an information-theoretic quantity, information-theoretic tools, such as the data-processing inequality and the chain rules of entropy [10], can be used to analyze the performance of machine learning algorithms. For example, we will illustrate this fact by presenting a simple proof to the classical generalization bound in the finite hypothesis space setting using, solely, information-theoretic inequalities without any reference to the union bound.

1.2. Types of Generalization

Generalization bounds can be stated either in expectation or in probability. Let $l : \mathcal{Z} \times \mathcal{H} \to [0,1]$ be some loss function with a bounded range. Then, we have the following definitions:

Definition 1 (Generalization in Expectation)**.** *The expected generalization risk of a learning algorithm* $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ *with respect to a loss* $l : \mathcal{Z} \times \mathcal{H} \to [0,1]$ *is defined by:*

$$R_{gen}(\mathcal{L}) = \mathbb{E}_{\boldsymbol{h}}[R(\boldsymbol{h})] - \mathbb{E}_{\boldsymbol{s},\boldsymbol{h}}\,\mathbb{E}_{\hat{\boldsymbol{z}}\sim\boldsymbol{s}}[l(\hat{\boldsymbol{z}},\boldsymbol{h})], \tag{2}$$

where $R(h)$ is defined in Equation (1), and the expectation is taken over the random choice of $\mathbf{s}$ and the internal randomness of $\mathcal{L}$. A learning algorithm $\mathcal{L}$ generalizes in expectation if $R_{gen}(\mathcal{L}) \to 0$ as $m \to \infty$ for all distributions $p(z)$.

Definition 2 (Generalization in Probability)**.** *A learning algorithm $\mathcal{L}$ generalizes in probability if for any $\epsilon > 0$, we have:*

$$p\Big\{\big|R(\boldsymbol{h}) - \mathbb{E}_{\hat{\boldsymbol{z}}\sim\boldsymbol{s}}[l(\hat{\boldsymbol{z}},\boldsymbol{h})]\big| > \epsilon\Big\} \to 0 \quad \text{as } m \to \infty,$$

where the probability is evaluated over the randomness of $\mathbf{s}$ and the internal randomness of the learning algorithm.

In general, both types of generalization have been used to analyze machine learning algorithms. For instance, generalization in probability is used in the VC theory to analyze algorithms with finite VC dimensions, such as linear classifiers [3]. Generalization in expectation, on the other hand, was used to analyze learning algorithms, such as the stochastic gradient descent (SGD), differential privacy, and ridge regression [11–14]. Generalization in expectation is often simpler to analyze, but it provides a weaker performance guarantee.

1.3. Paper Outline

In this paper, a third notion of generalization is introduced, which is called *uniform* generalization. Uniform generalization also provides generalization bounds in expectation, but it is stronger than the traditional form of generalization in expectation in Definition 1 because it requires that the generalization risk vanishes uniformly in expectation across *all* bounded parametric loss functions (hence the name). In this paper, a loss function $l : \mathcal{Z} \times \mathcal{H} \to [0, 1]$ is called " parametric" if it is conditionally independent of the original training sample given the learned hypothesis $h \in \mathcal{H}$.

As mentioned earlier, the *uniform* generalization risk is *equal* to an information-theoretic quantity and it yields classical results in statistical learning theory. Perhaps more importantly, and unlike traditional in-expectation guarantees that do not imply concentration, we will show that uniform generalization in expectation implies generalization in probability. Hence, all of the uniform generalization bounds derived in this paper hold both in expectation and with a high probability.

The theory of uniform generalization bridges the gap between information theory and statistical learning theory. For example, we will establish an equivalence relation between the VC dimension, on one hand, and another quantity that is quite analogous to the Shannon channel capacity, on the other hand. Needless to mention, both the VC dimension and the Shannon channel capacity are arguably the most central concepts in statistical learning theory and information theory. This connection between the two concepts is obtained via the notion of the " learning capacity" that we introduce in this paper, which is the supremum of the uniform generalization risk across all input distributions. We will compute the learning capacities for many machine learning algorithms and show how it matches known bounds on the generalization risk up to logarithmic factors.

In general, the main aim of this work is to bring to light a new information-theoretic approach for analyzing machine learning algorithms. Despite the fact that " uniform generalization" might appear to be a strong condition at a first sight, one of the central themes that is emphasized repeatedly throughout this paper is that uniform generalization is, in fact, a natural condition that arises commonly in practice. It is not a condition to require or enforce by machine learning practitioners! We believe this holds because any learning algorithm is a *channel* from the space of training samples to the hypothesis space so its risk for overfitting can be analyzed by studying the properties of this mapping itself. Such an approach yields the uniform generalization bounds that are derived in this paper.

While we strive to introduce foundational results in this work, there are many important questions that remain unanswered. We conclude this paper by listing some of those open problems and suggesting future directions of research.

2. Notation

The notation used in this paper is fairly standard. Important exceptions are listed here. If $\mathbf{x}$ is a random variable that takes its values from a finite set $\mathbf{s}$ uniformly at random, we write $\mathbf{x} \sim \mathbf{s}$ to denote such a distribution. If $\mathbf{x}$ is a boolean random variable (i.e., a predicate), then $\mathbb{I}\{\mathbf{x}\} = 1$ if and only if $\mathbf{x}$ is true, otherwise $\mathbb{I}\{\mathbf{x}\} = 0$. In general, random variables are denoted with boldface letters $\mathbf{x}$, instances of random variables are denoted with small letters x, matrices are denoted with capital letters X, and alphabets i.e., fixed sets) are denoted with calligraphic typeface $\mathcal{X}$ (except $\mathcal{L}$ that will be reserved for the learning algorithm and $\mathcal{D}$ that will be reserved for the input distribution as is customary in the literature).

Throughout this paper, we will always write $\mathcal{Z}$ to denote the space of observations (a.k.a. *domain*) and write $\mathcal{H}$ to denote the hypothesis space (a.k.a. *range*). A learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ is formally treated as a stochastic map, where the hypothesis $\mathbf{h} \in \mathcal{H}$ can be a deterministic or a randomized function of the training sample $\mathbf{s} \in \mathcal{Z}^m$. Given a 0–1 loss function $l : \mathcal{H} \times \mathcal{Z} \to \{0, 1\}$, we will abuse terminology slightly by speaking about the " VC dimension of $\mathcal{H}$" when we actually mean the VC dimension of the loss class $\{l(\cdot, h) : h \in \mathcal{H}\}$.

In addition, given two probability measures p and q defined on the same space, we will write $\langle p,\ q \rangle$ to denote the *overlapping coefficient* between p and q. That is, $\langle p,\ q \rangle = 1 - ||p\,,\ \ q||_{\mathcal{T}}$, where $||p\,,\ q||_{\mathcal{T}} = \frac{1}{2}||p - q||_1$ is the total variation distance.

Moreover, we will use the *order in probability* notation for real-valued *random* variables. Here, we adopt the notation used by [15] and [16]. In particular, let $\mathbf{x} = \mathbf{x}_n$ be a real-valued random variable that depends on some parameter $n \in \mathbb{N}$. Then, we will write $\mathbf{x}_n = O_p(f(n))$ if for any $\delta > 0$, there exists absolute constants C and n_0 such that for any fixed $n \geq n_0$, the inequality $|\mathbf{x}_n| < C\,|f(n)|$ holds with a probability of, at least, $1 - \delta$. In other words, the ratio $\mathbf{x}_n / f(n)$ is *stochastically* bounded [15]. Similarly, we write $\mathbf{x}_n = o_p(f(n))$ if $\mathbf{x}_n / f(n)$ converges to zero in probability. As an example, if $\mathbf{x} \sim \mathcal{N}(0, I_d)$ is a standard multivariate Gaussian vector, then $||\,\mathbf{x}\,||_2 = O_p(\sqrt{d})$ even though $||\,\mathbf{x}\,||_2$ can be arbitrarily large. Intuitively, the probability of the event $||\,\mathbf{x}\,||_2 \geq d^{\frac{1}{2}+\epsilon}$ when $\epsilon > 0$ goes to zero as $d \to \infty$ so $||\,\mathbf{x}\,||_2$ is *effectively* of the order $O(\sqrt{d})$.

3. Related Work

A learning algorithm is called *consistent* if the true risk of its hypothesis $\mathbf{h}$ converges to the optimal true risk in $\mathcal{H}$, i.e., $\inf_{h \in \mathcal{H}} R(h)$, as $m \to \infty$ in a distribution agnostic manner. A learning problem, which is a tuple $(\mathcal{Z},\ \mathcal{H},\ l)$ with l being a loss function defined on the product space $\mathcal{Z} \times \mathcal{H}$, is called *learnable* if it admits a consistent learning algorithm. It can be shown that learnability is equivalent to uniform convergence for supervised classification and regression even though uniform convergence is not necessary in the general setting [17].

Unlike learnability, the subject of generalization looks into how representative the empirical risk $R_{\mathbf{s}}(\mathbf{h})$ is to the true risk $R(\mathbf{h})$ as discussed earlier. It can be rightfully considered as an extension to the *law of large numbers*, which is one of the earliest and most important results in probability theory and statistics. However, unlike the law of large numbers, which assumes that observations are independent and identically distributed, the subject of generalization in machine learning addresses the case where the losses $l(\mathbf{z}_i, \mathbf{h})$ are no longer i.i.d. due to the fact that $\mathbf{h}$ is selected according to the training sample $\mathbf{s}$ and $\mathbf{z}_i \in \mathbf{s}$.

Similar to learnability, uniform convergence is, by definition, sufficient for generalization but it is not necessary because the learning algorithm might restrict its search space to a smaller subset of $\mathcal{H}$. So, in addition to uniform convergence bounds, several other methods have been introduced for bounding the generalization risk, such as using algorithmic stability, Rademacher and Gaussian complexities, generic chaining bounds, the PAC-Bayesian framework, and robustness-based analysis [5–7,18–20]. Classical concentration of measure inequalities, such as using the union bound, form the building blocks of such rich theories.

In this work, we address the subject of generalization in machine learning from an information-theoretic point of view. We will show that if the hypothesis $\mathbf{h}$ conveys " little" information about a random single training example $\hat{\mathbf{z}} \sim \mathbf{s}$, then the difference between $\mathbb{E}_{\hat{\mathbf{z}}\sim\mathbf{s}}[l(\hat{\mathbf{z}},\mathbf{h})]$ and $\mathbb{E}_{\mathbf{z}\sim p(z)}[l(\mathbf{z},\mathbf{h})]$ will be small with a high probability. The measure of information we use here is given by the notion of *variational information* $\mathcal{J}(\hat{\mathbf{z}};\mathbf{h})$ between the hypothesis $\mathbf{h}$ and a single random training example $\hat{\mathbf{z}} \sim \mathbf{s}$. Variational information, also sometimes called T-information [14], is an instance of the class of *informativity* measures using f-divergences, which can be motivated axiomatically [21,22]. Unlike traditional methods, we will prove that $\mathcal{J}(\hat{\mathbf{z}};\mathbf{h})$ is *equal* to the " uniform" generalization risk; it is not just an upper bound.

Information-theoretic approaches of analyzing the generalization risk of learning algorithms, such as the one proposed in this paper, have found applications in adaptive data analysis. This includes the work of [12] using the *max-information*, the work of [23] and [24] using the *mutual information*, and the work of [14] using the *leave-one-out* information. One key contribution of our work is to show that one should examine the relationship between the hypothesis and a *single* random training example, instead of examining the relationship between the hypothesis and the full training sample as is customary in the literature. The gap between such two approaches is strict. For example, Theorem 8 in Section 5.5 presents an example of when a learning algorithm can have a vanishing uniform generalization risk even when the mutual information between the learned hypothesis and the training sample can be made arbitrarily large.

4. Uniform Generalization

4.1. Preliminary Definitions

In this paper, we consider the general setting of learning introduced by Vapnik [3]. To reiterate, we have an observation space (a.k.a. domain) $\mathcal{Z}$ and a hypothesis space $\mathcal{H}$. Our learning algorithm $\mathcal{L}$ receives a set of m observations $\mathbf{s} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\} \in \mathcal{Z}^m$ generated i.i.d. from some fixed unknown distribution $p(z)$, and picks a hypothesis $\mathbf{h} \in \mathcal{H}$ according to some probability distribution $p(h \mid \mathbf{s})$. In other words, $\mathcal{L}$ is a channel from $\mathbf{s}$ to $\mathbf{h}$. In this paper, we allow the hypothesis $\mathbf{h}$ to be any *summary statistic* of the training set. It can be an answer to a query, a measure of central tendency, or a mapping from the input space to the output space. In fact, we even allow $\mathbf{h}$ to be a subset of the training set itself. In formal terms, $\mathcal{L}$ is a stochastic map between the two random variables $\mathbf{s} \in \mathcal{Z}^m$ and $\mathbf{h} \in \mathcal{H}$, where the exact interpretation of those random variables is irrelevant. Moreover, we assume that there exists a non-negative bounded loss function $l(z,h) \in [0,1]$ that is used to measure the fitness of the hypothesis $h \in \mathcal{H}$ on the observation $z \in \mathcal{Z}$.

For any fixed hypothesis $h \in \mathcal{H}$, we define its true risk $R(h)$ by Equation (1) and denote its empirical risk on the training sample by $R_{\mathbf{s}}(h)$. We also define the true and empirical risks of the *learning algorithm* $\mathcal{L}$ by the expected corresponding risk of its hypothesis:

$$R(\mathcal{L}) = \mathbb{E}_{\mathbf{s}}\,\mathbb{E}_{\mathbf{h}\sim p(h|\mathbf{s})}\left[R(\mathbf{h})\right] \quad = \mathbb{E}_{\mathbf{h}}\left[R(\mathbf{h})\right] \tag{3}$$

$$\hat{R}(\mathcal{L}) = \mathbb{E}_{\mathbf{s}}\,\mathbb{E}_{\mathbf{h}\sim p(h|\mathbf{s})}\left[R_{\mathbf{s}}(\mathbf{h})\right] \quad = \mathbb{E}_{\mathbf{s},\mathbf{h}}\left[R_{\mathbf{s}}(\mathbf{h})\right] \tag{4}$$

Finally, the generalization risk of the learning algorithm is defined by:

$$R_{gen}(\mathcal{L}) \doteq R(\mathcal{L}) - \hat{R}(\mathcal{L}) \tag{5}$$

Next, we define uniform generalization:

Definition 3 (Parametric Loss). *A loss function $l(\cdot, h) : \mathcal{Z} \to [0,1]$ is called parametric if it is conditionally independent of the training sample given the hypothesis $h \in \mathcal{H}$. That is, it satisfies the Markov chain $\boldsymbol{s} \to \boldsymbol{h} \to l(\cdot, \boldsymbol{h})$.*

Definition 4 (Uniform Generalization). *A learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ generalizes uniformly with rate $\epsilon \geq 0$ if for all bounded parametric losses $l : \mathcal{Z} \times \mathcal{H} \to [0,1]$, we have $|R_{gen}(\mathcal{L})| \leq \epsilon$, where $R_{gen}(\mathcal{L})$ is given in Equation (5).*

Informally, Definition 4 states that once a hypothesis **h** is selected by a learning algorithm $\mathcal{L}$ that achieves uniform generalization, then no " adversary" can post-process the hypothesis in a manner that causes over-fitting to occur. Equivalently, uniform generalization implies that the empirical performance of **h** on the sample **s** will remain close to its performance with respect to the underlying distribution regardless of how that performance is being measured. For example, the loss function $l : \mathcal{Z} \times \mathcal{H} \to [0,1]$ in Equation (5) can be the misclassification error rate as in the traditional classification setting, a cost-sensitive error rate as in fraud detection and medical diagnosis [25], or the Brier score as in probabilistic predictions [26]. The generalization guarantee would hold in any case.

4.2. Variational Information

Given two random variables **x** and **y**, the *variational information* between the two random variables is defined to be the total variation distance between the join distribution $p(\mathbf{x}, \mathbf{y})$ and the product of marginals $p(\mathbf{x}) \cdot p(\mathbf{y})$. We will denote this by $\mathcal{J}(\mathbf{x}; \mathbf{y})$. By definition:

$$\mathcal{J}(\mathbf{x}; \mathbf{y}) = \mathbb{E}_{\mathbf{x},\mathbf{y}} || p(\mathbf{x}, \mathbf{y}) \,,\ p(\mathbf{x}) \cdot p(\mathbf{y}) ||_{\mathcal{T}} = \mathbb{E}_{\mathbf{x}} || p(\mathbf{y}) \,,\ p(\mathbf{y}|\mathbf{x}) ||_{\mathcal{T}}$$

Note that $0 \leq \mathcal{J}(\mathbf{x}; \mathbf{y}) \leq 1$. We describe some of the important properties of variational information in this section. The reader may consult the appendices for detailed proofs.

Lemma 1 (Data Processing Inequality). *If $\boldsymbol{x} \to \boldsymbol{y} \to \boldsymbol{z}$ is a Markov chain, then:*

$$\mathcal{J}(\boldsymbol{x}; \boldsymbol{z}) \leq \mathcal{J}(\boldsymbol{y}; \boldsymbol{z})$$

This *data processing inequality* holds, in general, for all informativity measures using f-divergences [21,22].

Lemma 2 (Information Cannot Hurt). *For any random variables $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}$, and $\boldsymbol{z} \in \mathcal{Z}$, we have:*

$$\mathcal{J}(\boldsymbol{x}; \boldsymbol{y}) \leq \mathcal{J}(\boldsymbol{x}; (\boldsymbol{y}, \boldsymbol{z}))$$

Proof. The proof is in Appendix A. □

Finally, we derive a chain rule for the variational information.

Definition 5 (Conditional Variational Information)**.** *The conditional variational information between the two random variables* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *given* $\boldsymbol{z}$ *is defined by:*

$$\mathcal{J}(\boldsymbol{x};\, \boldsymbol{y} \,|\, \boldsymbol{z}) = \mathbb{E}_{\boldsymbol{z}}\big[||p(\boldsymbol{x}, \boldsymbol{y} \,|\, \boldsymbol{z}),\ \ p(\boldsymbol{x}|\boldsymbol{z}) \cdot p(\boldsymbol{y}|\boldsymbol{z})||_{\mathcal{T}}\big],$$

which is analogous to the conditional mutual information in information theory [10].

Theorem 1 (Chain Rule)**.** *Let* $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k)$ *be a sequence of random variables. Then, for any random variable* $\boldsymbol{z}$*, we have:* $\mathcal{J}(\boldsymbol{z};\, (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k)) \leq \sum_{t=1}^{k} \mathcal{J}(\boldsymbol{z};\, \boldsymbol{h}_t \,|\, (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{t-1}))$

Proof. The proof is in Appendix B. □

Although the chain rule above provides an upper bound, the upper bound is tight in the following sense:

Proposition 1. *For any random variables* $\boldsymbol{x}, \boldsymbol{y}$*, and* $\boldsymbol{z}$*, we have* $\big|\mathcal{J}(\boldsymbol{x};\, (\boldsymbol{y}, \boldsymbol{z})) - \mathcal{J}(\boldsymbol{x};\, \boldsymbol{z} \,|\, \boldsymbol{y})\big| \leq \mathcal{J}(\boldsymbol{x};\, \boldsymbol{y})$ *and* $\big|\mathcal{J}(\boldsymbol{x};\, (\boldsymbol{y}, \boldsymbol{z})) - \mathcal{J}(\boldsymbol{x};\, \boldsymbol{y})\big| \leq \mathcal{J}(\boldsymbol{x};\, \boldsymbol{z} \,|\, \boldsymbol{y})$*.*

Proof. The proof is in Appendix C. □

In other words, the inequality in the chain rule $\mathcal{J}(\mathbf{x};\, (\mathbf{y}, \mathbf{z})) \leq \mathcal{J}(\mathbf{x};\, \mathbf{y}) + \mathcal{J}(\mathbf{x};\, \mathbf{z} \,|\, \mathbf{y})$ becomes an equality if:

$$\min\{\mathcal{J}(\mathbf{x};\, \mathbf{y}),\ \mathcal{J}(\mathbf{x};\, \mathbf{z} \,|\, \mathbf{y})\} = 0$$

The chain rule provides a recipe for computing the bias of a composition of hypotheses $(\mathbf{h}_1, \ldots, \mathbf{h}_k)$. Recently, [23] proposed an *information budget* framework for controlling the bias of estimators by controlling the mutual information between $\mathbf{h}$ and the training sample $\mathbf{s}$. The proposed framework rests on the chain rule of mutual information. Here, we note that the argument for the information budget framework also holds when using the variational information due to the chain rule above.

4.3. Equivalence Result

Our first main theorem states that the uniform generalization risk has a precise information-theoretic characterization.

Theorem 2. *Given a fixed constant* $0 \leq \epsilon \leq 1$ *and a learning algorithm* $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ *that selects a hypothesis* $\boldsymbol{h} \in \mathcal{H}$ *according to a training sample* $\boldsymbol{s} = \{z_1, \ldots, z_m\}$*, where* $z_i \sim p(z)$ *are i.i.d.,* $\mathcal{L}$ *generalizes uniformly with rate* ϵ *if and only if* $\mathcal{J}(\boldsymbol{h};\, \hat{\boldsymbol{z}}) \leq \epsilon$*, where* $\hat{\boldsymbol{z}} \sim \boldsymbol{s}$ *is a single random training example.*

Proof. Let $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ be a learning algorithm that receives a finite set of training examples $\mathbf{s} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\} \in \mathcal{Z}^m$ drawn i.i.d. from a fixed unknown distribution $p(z)$. Let $\mathbf{h} \sim p(h|\mathbf{s})$ be the hypothesis chosen by $\mathcal{L}$ (can be deterministic or randomized) and write $\hat{\mathbf{z}} \sim \mathbf{s}$ to denote a random variable that selects its value uniformly at random from the training sample $\mathbf{s}$. Clearly, $\hat{\mathbf{z}}$ and $\mathbf{h}$ are not independent in general. To simplify notation, we will write $\mathbf{l} = l(\cdot, \mathbf{h}) : \mathcal{Z} \to [0, 1]$ to denote the loss function. Note that $\mathbf{l}$ is itself a random variable that satisfies the Markov chain $\mathbf{s} \to \mathbf{h} \to \mathbf{l}$. The claim is that $\mathcal{L}$ generalizes uniformly with rate $\epsilon > 0$ across all parametric loss functions $\mathbf{l}$ if and only if $\mathcal{J}(\mathbf{h};\, \hat{\mathbf{z}}) \leq \epsilon$.

By the Markov property, we have $p(\mathbf{l}|\mathbf{h},\mathbf{s}) = p(\mathbf{l}|\mathbf{h})$. By definition, the true and empirical risks of $\mathcal{L}$ are given by:

$$R(\mathcal{L}) = \mathbb{E}_{\mathbf{s},\mathbf{h}}\, \mathbb{E}_{\mathbf{l}|\mathbf{h}} \mathbb{E}_{\mathbf{z}\sim p(z)}\, \mathbf{l}(\mathbf{z}) = \mathbb{E}_{\mathbf{l}}\, \mathbb{E}_{\mathbf{z}\sim p(z)}\, \mathbf{l}(\mathbf{z}) \tag{6}$$

$$\hat{R}(\mathcal{L}) = \mathbb{E}_{\mathbf{s}}\, \mathbb{E}_{\mathbf{l}|\mathbf{s}}\, \mathbb{E}_{\mathbf{z}\sim \mathbf{s}}\, \mathbf{l}(\mathbf{z}) = \mathbb{E}_{\mathbf{l}}\, \mathbb{E}_{\mathbf{s}|\mathbf{l}}\, \mathbb{E}_{\mathbf{z}\sim \mathbf{s}}\, \mathbf{l}(\mathbf{z}) \tag{7}$$

Because $\hat{\mathbf{z}} \sim \mathbf{s}$ is a random variable whose value is chosen uniformly at random with replacement from the training set $\mathbf{s}$, its marginal distribution is $p(z)$. Its *conditional* distribution given $\mathbf{l}$ can be different, however, because both $\mathbf{l}$ and $\hat{\mathbf{z}}$ depend on the training set $\mathbf{s}$. However, they are both *conditionally* independent of each other given $\mathbf{s}$. By marginalization, we have:

$$p(\hat{\mathbf{z}}|\mathbf{l}) = \mathbb{E}_{\mathbf{s}|\mathbf{l}}\, p(\hat{\mathbf{z}}|\mathbf{s},\mathbf{l}) = \mathbb{E}_{\mathbf{s}|\mathbf{l}}\, p(\hat{\mathbf{z}}|\mathbf{s})$$

Combining this with Equations (6) and (7) yields $R(\mathcal{L}) = \mathbb{E}_{\mathbf{l}}\, \mathbb{E}_{\hat{\mathbf{z}}}\, \mathbf{l}(\hat{\mathbf{z}})$ and $\hat{R}(\mathcal{L}) = \mathbb{E}_{\mathbf{l}}\, \mathbb{E}_{\hat{\mathbf{z}}|\mathbf{l}}\, \mathbf{l}(\hat{\mathbf{z}})$. Both equations imply that:

$$R(\mathcal{L}) - \hat{R}(\mathcal{L}) = \mathbb{E}_{\mathbf{l}}\left[\mathbb{E}_{\hat{\mathbf{z}}}\, \mathbf{l}(\hat{\mathbf{z}}) - \mathbb{E}_{\hat{\mathbf{z}}|\mathbf{l}}\, \mathbf{l}(\hat{\mathbf{z}})\right]$$

Now, we would like to sandwich the right-hand side between upper and lower bounds. To do this, we note that if $p_1(z)$ and $p_2(z)$ are two distributions defined on the same domain $\mathcal{Z}$ and $f : \mathcal{Z} \to [0,1]$, then:

$$\left|\mathbb{E}_{\mathbf{z}\sim p_1(z)}\, f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}\sim p_2(z)}\, f(\mathbf{z})\right| \leq ||p_1(z)\,,\ p_2(z)||_{\mathcal{T}},$$

where $||p_1(z)\,,\ p_2(z)||_{\mathcal{T}}$ is the total variation distance. This result can be immediately proven by considering the two regions $\{z \in \mathcal{Z} :\ p_1(z) > p_2(z)\}$ and $\{z \in \mathcal{Z} :\ p_1(z) < p_2(z)\}$ separately. In addition, it is tight because the inequality holds with equality for the loss function $f(z) = \mathbb{I}\{p_1(z) \geq p_2(z)\}$. Consequently:

$$|R(\mathcal{L}) - \hat{R}(\mathcal{L})| \leq \mathcal{J}(\mathbf{l};\, \hat{\mathbf{z}})$$

Finally, from the Markov chain $\hat{\mathbf{z}} \to \mathbf{s} \to \mathbf{h} \to \mathbf{l}$ and the data processing inequality, we have $\mathcal{J}(\mathbf{l};\, \hat{\mathbf{z}}) \leq \mathcal{J}(\mathbf{h};\, \hat{\mathbf{z}})$. Plugging this into the earlier inequality yields the bound:

$$|R(\mathcal{L}) - \hat{R}(\mathcal{L})| \leq \mathcal{J}(\mathbf{h};\, \hat{\mathbf{z}})$$

To prove the converse, define:

$$\begin{aligned} l^{\star}(z,\mathbf{h}) &= \mathbb{I}\big\{p(\hat{\mathbf{z}} = z)\ \geq p(\hat{\mathbf{z}} = z\,|\,\mathbf{h})\big\} \\ &= \mathbb{I}\big\{p(\hat{\mathbf{z}} = z)\ \geq\ \mathbb{E}_{\mathbf{s}\,|\,\mathbf{h}}\ [p_{\hat{\mathbf{z}}\sim\mathbf{s}}\ (\hat{\mathbf{z}} = z)]\big\} \end{aligned}$$

The loss $l^{\star}(z,\mathbf{h})$ is independent of the training sample given $\mathbf{h}$ because $p(\hat{\mathbf{z}} = z\,|\,\mathbf{h})$ is evaluated by taking expectation over all the training samples conditioned on $\mathbf{h}$. Hence, $l^{\star}(z,\mathbf{h})$ is a 0–1 loss defined on the product space $\mathcal{Z} \times \mathcal{H}$ and satisfies the Markov chain $\mathbf{s} \to \mathbf{h} \to \mathbf{l}$. However, given this choice of loss, we have:

$$\begin{aligned} |R(\mathcal{L}) - \hat{R}(\mathcal{L})| &= \mathbb{E}_{\mathbf{h}}\big[\mathbb{E}_{\hat{\mathbf{z}}}\ \mathbb{I}\{p(\hat{\mathbf{z}}) > p(\hat{\mathbf{z}}\,|\,\mathbf{h})\} - \mathbb{E}_{\hat{\mathbf{z}}\,|\,\mathbf{h}}\ \mathbb{I}\{p(\hat{\mathbf{z}}) > p(\hat{\mathbf{z}}\,|\,\mathbf{h})\}\big] \\ &= \mathbb{E}_{\mathbf{h}}||p(\hat{\mathbf{z}})\,,\ p(\hat{\mathbf{z}}\,|\,\mathbf{h})||_{\mathcal{T}} = \mathcal{J}(\mathbf{h};\, \hat{\mathbf{z}}) \end{aligned}$$

Hence, the variational information $\mathcal{J}(\mathbf{h};\, \hat{\mathbf{z}})$ does not only provide an upper bound on the uniform generalization risk, but is also a lower bound to it. Therefore, $\mathcal{J}(\mathbf{h};\, \hat{\mathbf{z}})$ is equal to the uniform generalization risk. □

Remark 1. *One important observation about Theorem 2 is that the variational information is measured between the hypothesis $\boldsymbol{h}$ and a* single *training example $\hat{\boldsymbol{z}}$, which is quite different from previous works that looked into the mutual information with the entire training sample $\boldsymbol{s}$. By considering $\hat{\boldsymbol{z}}$ rather than $\boldsymbol{s}$, we quantify the uniform generalization risk with equality and the resulting bound is not vacuous even if the learning algorithm was deterministic. By contrast, $\mathcal{J}(\boldsymbol{s};\boldsymbol{h})$ may yield vacuous bounds when $\mathcal{L}$ is deterministic and both $\mathcal{Z}$ and $\mathcal{H}$ are uncountable.*

For concreteness, we illustrate how to compute the uniform generalization risk (or equivalently the variational information) on two simple examples. Here, $B(k;\phi,n) = \binom{n}{k}\phi^k\,(1-\phi)^{n-k}$ is the binomial distribution. The first example is a special case of a more general theorem that will be presented later in Section 5.2.

Example 1. *Suppose that observations $\boldsymbol{z}_i \in \{0,1\}$ are i.i.d. Bernoulli trials with $p(\boldsymbol{z}_i = 1) = \phi$, and that the hypothesis produced by $\mathcal{L}$ is the empirical average $\boldsymbol{h} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{z}_i$. Because $p(\boldsymbol{h} = k/m \,\big|\, \boldsymbol{z}_{trn} = 1) = B(k-1;\phi,m-1)$ and $p(\boldsymbol{h} = k/m \,\big|\, \boldsymbol{z}_{trn} = 0) = B(k;\phi,m-1)$, it can be shown that the uniform generalization risk of this learning algorithm is given by the following quantity assuming that ϕm is an integer:*

$$\mathcal{J}(\hat{\boldsymbol{z}};\boldsymbol{h}) = 2\,(1-\phi)^{(1-\phi)\,m}\;\phi^{1+m\,\phi}\,(1+m\,\phi)\,\binom{m}{m\,\phi+1} \tag{8}$$

This is maximized when $\phi = 1/2$, in which case, the uniform generalization risk can be bounded using the Stirling approximation [27] by $1/\sqrt{2\,\pi\,m}$ up to a first-order term.

Proof. First, the probability we obtain a hypothesis $\mathbf{h} = \frac{k}{m}$, where $k \in \{0,1,\ldots,m\}$, given that we have m Bernoulli trials has a binomial distribution:

$$p(\mathbf{h} = \frac{k}{m}) = \binom{m}{k}\,\phi^k\,(1-\phi)^{m-k}$$

We use the identity:

$$\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{h}) = \sum_{k=0}^{m}\; p\big(\mathbf{h} = \frac{k}{m}\big)\;||p(\hat{\mathbf{z}})\,,\;\; p(\hat{\mathbf{z}}|\mathbf{h})||_{\mathcal{T}}$$

However, $p(\hat{\mathbf{z}})$ is Bernoulli with probability of success ϕ while $p(\hat{\mathbf{z}}|\mathbf{h} = \frac{k}{m})$ is Bernoulli with probability of success $\mathbf{h}$. The total variation distance between the two Bernoulli distributions is given by $|\phi - \mathbf{h}|$. So, we obtain:

$$\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{h}) = \sum_{k=0}^{m}\binom{m}{k}\,\phi^k\,(1-\phi)^{m-k}\,\Big|\phi - \frac{k}{m}\Big| \tag{9}$$

This is the *mean deviation*. Assuming $\phi\,m$ is an integer, then the mean deviation of the binomial random variable is given by de Moivre's formula:

$$MD = 2\,(1-\phi)^{(1-\phi)\,m}\;\phi^{1+m\,\phi}\,(1+m\,\phi)\,\binom{m}{m\,\phi+1} \tag{10}$$

The mean deviation is maximized when $\phi = \frac{1}{2}$. This gives us:

$$\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{h}) \leq \frac{1}{2^m}\binom{m}{m/2+1} \sim \frac{1}{\sqrt{2\,\pi\,m}},$$

where in the last step we expanded the binomial coefficient and used Stirling's approximation [27]. □

Example 2. *Suppose that the domain is $\mathcal{Z} = \{1, 2, 3, \ldots, K\}$ for some $K < \infty$, where $p(\boldsymbol{z} = k) = 1/K$ for all $k \in \mathcal{Z}$. Let the hypothesis space be $\mathcal{H} = \mathcal{Z}$ where $p(\boldsymbol{h} = k)$ is equal to the fraction of times the value k is observed in the training sample $\boldsymbol{s} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$. For example, if $\boldsymbol{s} = \{1, 3, 2, 1, 1, 3\}$, the hypothesis $\boldsymbol{h}$ is chosen among the set $\{1, 2, 3\}$ with the respective probabilities $\{1/2, 1/6, 1/3\}$. Then, the variational information is given by:*

$$\mathcal{J}(\hat{\boldsymbol{z}}; \boldsymbol{h}) = \frac{1}{m}\Big(1 - \frac{1}{K}\Big)$$

Proof. We have by symmetry $p(\mathbf{h} = k) = 1/K$ for all $k \in \{1, 2, 3, \ldots, K\}$. Let $\hat{\mathbf{z}} = x$. By Bayes rule, we have:

$$\begin{aligned} p(\hat{\mathbf{z}} = x | \mathbf{h} = k) &= p(\mathbf{h} = k | \hat{\mathbf{z}} = x) \cdot \frac{p(\hat{\mathbf{z}} = x)}{p(\mathbf{h}) = k} \\ &= p(\mathbf{h} = k | \hat{\mathbf{z}} = x) \end{aligned}$$

However, given one observation $\hat{\mathbf{z}} = x$, the probability of selecting a hypothesis $\mathbf{h} = k$ depends on two cases:

$$p(\mathbf{h} = k \,|\, \hat{\mathbf{z}} = x) = \begin{cases} q & \text{if } k = x \\ r & \text{if } k \neq x \end{cases}$$

for some values $q \geq 0$ and $r \geq 0$ such that $q + (K - 1)\, r = 1$. To find q, we use the definition of $\mathcal{L}$:

$$q = \frac{1}{m} + \frac{1}{K} \cdot \frac{m - 1}{m} = \frac{1}{K} + \frac{1}{m}\Big(1 - \frac{1}{K}\Big)$$

This holds because $\mathcal{L}$ is equivalent to an algorithm that selects a single observation in the set $\mathbf{s}$ uniformly at random. So, to satisfy the condition $q + (K - 1)\, r = 1$, we have:

$$r = \frac{1}{K} - \frac{1}{mK}$$

Now, we are ready to find the desired expression.

$$\begin{aligned} \mathcal{J}(\hat{\mathbf{z}}; \mathbf{h}) &= \frac{1}{2} \sum_{x \in \mathcal{Z}} p(\hat{\mathbf{z}} = x) \sum_{k \in \mathcal{Z}} \big| p(\mathbf{h} = k) - p(\mathbf{h} = k | \hat{\mathbf{z}} = x) \big| \\ &= \frac{1}{2} \sum_{k \in \mathcal{Z}} \big| p(\mathbf{h} = k) - p(\mathbf{h} = k | \hat{\mathbf{z}} = 1) \big| \\ &= \frac{1}{2}\Big[\frac{1}{m}\big(1 - \frac{1}{K}\big) + \frac{K - 1}{mK}\Big] = \frac{1}{m}\Big(1 - \frac{1}{K}\Big) \quad \square \end{aligned}$$

Note that the variational information in Example 2 is $\Theta(1/m)$, which is smaller than the variational information in Example 1. This is not a coincidence. The difference between the two examples is related to *data processing*. Specifically, suppose that $K = 2$ in Example 2 and let $\mathbf{h}_2$ be the hypothesis. Let $\mathbf{h}_1$ be the hypothesis in Example 1. Then, we have the Markov chain $\mathbf{s} \to \mathbf{h}_1 \to \mathbf{h}_2$ because $\mathbf{h}_2$ is Bernoulli with parameter $\mathbf{h}_1$.

4.4. Learning Capacity

The variational information depends on the distribution of observations $p(z)$, which is seldom known in practice. To construct a distribution-free bound on the uniform generalization risk, we introduce the following quantity:

Definition 6 (Learning Capacity)**.** *The learning capacity of an algorithm $\mathcal{L}$ is defined by:*

$$C(\mathcal{L}) \doteq \sup_{p(z)} \{\mathcal{J}(\hat{\boldsymbol{z}}; \boldsymbol{h})\}, \tag{11}$$

where $\boldsymbol{h}$ and $\hat{\boldsymbol{z}}$ are as defined in Theorem 2.

The above quantity is analogous to the Shannon channel capacity except that it is measured in the total variation distance. It quantifies the capacity for overfitting in the given learning algorithm. For example, the learning capacity of the algorithm in Example 1 is $1/\sqrt{2\pi m}$ up to a first order term, as proved earlier, so its capacity for overfitting is larger than that of the learning algorithm in Example 2.

Theorem 2 reveals that $C(\mathcal{L})$ has, at least, three *equivalent* interpretations:

1. *Statistical*: The learning capacity $C(\mathcal{L})$ is equal to the supremum of the expected generalization risk $R_{gen}(\mathcal{L})$ across all input distributions and all bounded parametric losses. This holds by Theorem 2 and Definition 6.
2. *Information-Theoretic*: The learning capacity $C(\mathcal{L})$ is equal to the amount of information contained in the hypothesis $\mathbf{h}$ about the training examples. This holds because $\mathcal{J}(\hat{\mathbf{z}}; \mathbf{h}) = \mathbb{E}_{\mathbf{h}}||p(\hat{\mathbf{z}})\,,\ p(\hat{\mathbf{z}}\,|\,\mathbf{h})||_{\mathcal{T}}$.
3. *Algorithmic*: The learning capacity $C(\mathcal{L})$ measures the influence of a single training example $\hat{\mathbf{z}}$ on the distribution of the final hypothesis $\mathbf{h}$. As such, a learning algorithm has a small learning capacity if and only if it is algorithmically stable. This follows from the fact that $\mathcal{J}(\hat{\mathbf{z}}; \mathbf{h}) = \mathbb{E}_{\hat{\mathbf{z}}}||p(\mathbf{h})\,,\ p(\mathbf{h}\,|\,\hat{\mathbf{z}})||_{\mathcal{T}}$.

Throughout the sequel, we analyze the properties of $C(\mathcal{L})$ and derive upper bounds for it under various conditions, such as in the finite hypothesis space setting and differential privacy.

4.5. The Definition of Hypothesis

In the proof of Theorem 2, the following Markov chain $\hat{\mathbf{z}} \to \mathbf{s} \to \mathbf{h} \to \mathbf{l}(\cdot, \mathbf{h})$ is used. Essentially, this states that the loss function $\mathbf{l}(\cdot, \mathbf{h}) : \mathcal{Z} \to [0, 1]$, which is a random variable itself, must be parameterized entirely by the hypothesis $\mathbf{h}$ as stated in Definition 3. We list, next, a few examples that highlight this point.

Example 3 (Input Normalization)**.** *If the data is normalized prior to training, such as using min-max or z-score normalization, then the normalization parameters are included in the definition of the hypothesis $\boldsymbol{h}$.*

Example 4 (Feature Selection)**.** *If the observations $\boldsymbol{z}$ comprise of d features and feature selection is implemented prior to training a model $\boldsymbol{v}$ (such as in classification or clustering), then the hypothesis $\boldsymbol{h}$ is the composition $(\boldsymbol{u}, \boldsymbol{v})$, where $\boldsymbol{u} \in \{0, 1\}^d$ encodes the set of the features that have been selected by the feature selection algorithm.*

Example 5 (Cross Validation)**.** *Hyper-parameter tuning is a common practice in machine learning. This includes choosing the tradeoff parameter C in support vector machine (SVM) [28] or the bandwidth γ in radial basis function (RBF) networks [29]. However, not all hyper-parameters are encoded in the hypothesis $\boldsymbol{h}$. For instance, the tradeoff constant C is never used during prediction so it is omitted from the definition of $\boldsymbol{h}$ but the bandwidth parameter γ is included if it is selected based on the training sample.*

In order to illustrate why the Markov chain $\hat{\mathbf{z}} \to \mathbf{s} \to \mathbf{h} \to \mathbf{l}(\cdot, \mathbf{h})$ is important, consider the following simple scenario. Suppose we have a mixture of two Gaussians in $\mathbb{R}^d$, one corresponding to the positive class and one corresponding to the negative class. If z-score normalization is applied before training a linear classifier, then the generalization risk might increase with normalization because the final hypothesis now includes more information about the training sample (see Lemma 2). Figure 1 shows this effect when

$d = 1$. As illustrated in the figure, normalization is often important in order to assign equal weights to all features but it can increase the generalization risk as well.

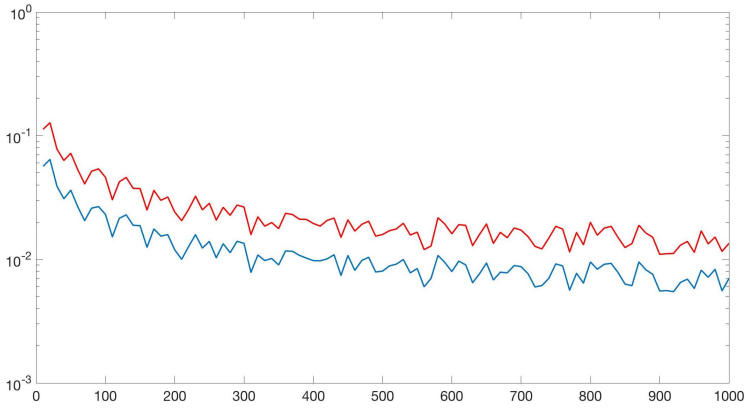

Figure 1. This figure corresponds to a classification problem in one dimension in which a classifier is a threshold between positive and negative examples. In this figure, the x axis is the number of training examples while the y-axis is the generalization risk. The red curve (top) corresponds to the difference between training and test accuracy when z-score normalization is applied before learning a classifier. The blue curve (bottom) corresponds to the difference between training and test accuracy when the data is not normalized.

4.6. Concentration

The notion of uniform generalization in Definition 4 provides *in-expectation* guarantees. In this section, we show that whereas traditional generalization in expectation does not imply concentration, *uniform* generalization in expectation implies concentration. In fact, we will use the chain rule in Theorem 1 to derive a Markov-type inequality. After that, we show that the bound is tight.

We begin by showing why a non-uniform generalization in expectation does not imply concentration.

Proposition 2. *There exists a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ and a parametric loss $l : \mathcal{Z} \times \mathcal{H} \to [0, 1]$ such that the expected generalization risk is $R_{gen}(\mathcal{L}) = 0$ even though $p\{|R(\boldsymbol{h}) - R_{\boldsymbol{s}}(\boldsymbol{h})| = \frac{1}{2}\} = 1$, where the probability is evaluated over the randomness of $\boldsymbol{s}$ and the internal randomness of $\mathcal{L}$.*

Proof. Let $\mathcal{Z} = [0, 1]$ be an instance space with a continuous marginal density $p(z)$ and let $\mathcal{Y} = \{-1, +1\}$ be the target set. Let $h^\star : \mathcal{Z} \to \{-1, +1\}$ be some *fixed* predictor, such that $p\{h^\star(\mathbf{z}) = 1\} = \frac{1}{2}$, where the probability is evaluated over the random choice of $\mathbf{z} \in \mathcal{Z}$. In other words, the marginal distribution of the labels predicted by $h^\star$ is uniform over the set $\{-1, +1\}$. These assumptions are satisfied, for example, if $p(z)$ is uniform in $[0, 1]$ and $h^\star(z) = \mathbb{I}\{z < 1/2\}$.

Next, let the hypothesis space $\mathcal{H}$ be the set of predictors from $\mathcal{Z}$ to $\{-1, +1\}$ that output a label in $\{-1, +1\}$ uniformly at random everywhere in $\mathcal{Z}$ except at a finite number of points. Define the parametric loss by $l(z; h) = \mathbb{I}\{h(z) \neq h^\star(z)\}$.

Next, we construct a learning algorithm $\mathcal{L}$ that generalizes perfectly in expectation but does not generalize in probability. The learning algorithm $\mathcal{L}$ simply picks $\mathbf{h} \in \{\mathbf{h}_0, \mathbf{h}_1\}$ at random with equal probability. The two hypotheses are:

$$\mathbf{h}_0(z) = \begin{cases} -h^\star(z) & \text{if } z \in \mathbf{s} \\ \text{Uniform}(-1,+1) & \text{if } z \notin \mathbf{s} \end{cases}$$

$$\mathbf{h}_1(z) = \begin{cases} h^\star(z) & \text{if } z \in \mathbf{s} \\ \text{Uniform}(-1,+1) & \text{if } z \notin \mathbf{s} \end{cases}$$

Because $\mathcal{Z}$ is uncountable, where the probability of seeing the same observation $\mathbf{z}$ twice is zero, $R(\mathbf{h}) = \frac{1}{2}$ for this learning algorithm. Thus:

$$R_{gen}(\mathcal{L}) = \mathbb{E}_{\mathbf{s},\mathbf{h}}\big[R_{\mathbf{s}}(\mathbf{h}) - R(\mathbf{h})\big] = 0$$

However, the empirical risk for any $\mathbf{s}$ satisfies $R_{\mathbf{s}}(\mathbf{h}) \in \{0,1\}$ while the true risk always satisfies $R(\mathbf{h}) = \frac{1}{2}$, as mentioned earlier. Hence, the statement of the proposition follows. □

There are many ways of seeing why the algorithm in Proposition 2 does not generalize *uniformly* in expectation. The simplest way is to use the equivalence between uniform generalization and variational information as stated in Theorem 2. Given the hypothesis $\mathbf{h} \in \{\mathbf{h}_0, \mathbf{h}_1\}$ that is learned by the algorithm constructed in the proposition, the marginal distribution of an individual training example $p(\hat{\mathbf{z}} \,|\, \mathbf{h})$ is uniform over the sample $\mathbf{s}$. This follows from the fact that the hypothesis $\mathbf{h}$ has to encode the entire sample $\mathbf{s}$. However, the probability of seeing the same observation twice is zero (by construction). Hence, $||p(\hat{\mathbf{z}})\,,\ p(\hat{\mathbf{z}} \mid \mathbf{h})||_{\mathcal{T}} = 1$. This shows that $C(\mathcal{L}) = 1$.

The example in Proposition 2 reveals an interesting property of non-uniform generalization. Namely, *non-uniform* generalization can be sensitive to every bit of information provided by the hypothesis. In the example above, the hypothesis $\mathbf{h}$ is encoded by the pair $(\mathbf{s}, \mathbf{k})$, where $\mathbf{k} \in \{0,1\}$ determines which of the two hypotheses $\{\mathbf{h}_0, \mathbf{h}_1\}$ is selected. The discrepancy between generalization in expectation and generalization in probability happens because $\mathbf{k}$ is added into the hypothesis.

Next, we use the chain rule in Theorem 1 to prove that uniform generalization, on the other hand, is a *robust* property of learning algorithms. More precisely, if $\mathbf{k}$ has a finite domain, then a hypothesis $\mathbf{h}$ generalizes uniformly in expectation if and only if the pair $(\mathbf{h}, \mathbf{k})$ generalizes uniformly in expectation. Hence, adding any finite amount of information (in bits) to a hypothesis cannot alter its uniform generalization property in a significant way.

Theorem 3. *Let $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ be a learning algorithm whose hypothesis is $\boldsymbol{h} \in \mathcal{H}$. Let $\boldsymbol{k} \in \mathcal{K}$ be a different hypothesis that is obtained from the same sample $\boldsymbol{s}$. If $\hat{\boldsymbol{z}} \sim \boldsymbol{s}$, then:*

$$\mathcal{J}(\hat{\boldsymbol{z}};\, (\boldsymbol{h}, \boldsymbol{k})) \leq (2 + \frac{|\mathcal{K}|}{2}) \cdot \mathcal{J}(\hat{\boldsymbol{z}};\, \boldsymbol{h}) + \sqrt{\frac{\log |\mathcal{K}|}{2m}}$$

Proof. The proof is in Appendix D. □

We use Theorem 3, next, to prove that a uniform generalization in expectation implies a generalization in probability. The proof is by contradiction. Suppose we have a hypothesis $\mathbf{h}$ that generalizes uniformly in expectation but there exists a parametric loss $l : \mathcal{Z} \times \mathcal{H} \to [0,1]$ that does not generalize in probability. We will derive a contradiction from these two assumptions. We show that appending little information to the hypothesis $\mathbf{h}$ will allow us to construct a *different* parametric loss that does not generalize in expectation

by determining whether or not the empirical risk w.r.t. $l : \mathcal{Z} \times \mathcal{H} \to [0,1]$ is greater than, approximately equal to, or is less than the true risk w.r.t. the same loss. This is described in, at most, two bits. Knowing this additional information, we can define a new parametric loss that does not generalize in expectation, which contradicts the definition of uniform generalization.

Theorem 4. *Let $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ be a learning algorithm, whose risk is evaluated using a parametric loss $l : \mathcal{Z} \times \mathcal{H} \to [0,1]$. Then:*

$$p\Big\{\big|R_{\mathbf{s}}(\boldsymbol{h}) - R(\boldsymbol{h})\big| \geq t\Big\} \leq \frac{7}{2t}\Big[\mathcal{J}(\hat{\boldsymbol{z}}; \boldsymbol{h}) + \sqrt{\frac{\log 3}{49m}}\Big],$$

where the probability is evaluated over the random choice of $\mathbf{s}$ and the internal randomness of $\mathcal{L}$.

Proof. Let $l : \mathcal{Z} \times \mathcal{H} \to [0,1]$ be a parametric loss function and write:

$$\kappa(t) = p\Big\{\big|R_{\mathbf{s}}(\mathbf{h}) - R(\mathbf{h})\big| \geq t\Big\} \tag{12}$$

Consider the new pair of hypotheses $(\mathbf{h}, \mathbf{k})$, where:

$$\mathbf{k} = \begin{cases} +1, & \text{if } R_{\mathbf{s}}(\mathbf{h}) \geq R(\mathbf{h}) + t \\ -1, & \text{if } R_{\mathbf{s}}(\mathbf{h}) \leq R(\mathbf{h}) - t \\ 0, & \text{otherwise} \end{cases}$$

Then, by Theorem 3, the uniform generalization risk in expectation for the composition of hypotheses $(\mathbf{h}, \mathbf{k})$ is bounded by $(7/2)\,\mathcal{J}(\hat{\mathbf{z}}; \mathbf{h}) + \sqrt{\frac{\log 3}{2m}}$. This holds uniformly across all parametric loss functions that satisfy the Markov chain $\mathbf{s} \to (\mathbf{h}, \mathbf{k}) \to \mathbf{l}(\cdot, (\mathbf{h}, \mathbf{k}))$. Next, consider the parametric loss:

$$\mathbf{l}(z, (\mathbf{h}, \mathbf{k})) = \begin{cases} l(z; \mathbf{h}) & \text{if } \mathbf{k} = +1 \\ 1 - l(z; \mathbf{h}) & \text{if } \mathbf{k} = -1 \\ 0 & \text{otherwise} \end{cases}$$

Note that $\mathbf{l}(z, (\mathbf{h}, \mathbf{k}))$ is parametric with respect to the composition of hypotheses $(\mathbf{h}, \mathbf{k})$. Using Equation (12), the generalization risk w.r.t $\mathbf{l}(z, (\mathbf{h}, \mathbf{k}))$ in expectation is, at least, as large as $t\,\kappa(t)$. Therefore, by Theorems 2 and 3, we have $t\,\kappa(t) \leq (7/2)\,\mathcal{J}(\hat{\mathbf{z}}; \mathbf{h}) + \sqrt{\frac{\log 3}{2m}}$, which is the statement of the theorem (Note: The proof assumes that the loss function $\mathbf{l}$ has access to the underlying distribution. This assumption is valid because the underlying distribution $p(z)$ is fixed and does not depend on any random outcomes, such as $\mathbf{s}$ or $\mathbf{h}$). □

Theorem 4 reveals that uniform generalization is sufficient for concentration to hold. Importantly, the generalization bound depends on the learning algorithm $\mathcal{L}$ only via its variational information $\mathcal{J}(\hat{\mathbf{z}}; \mathbf{h})$. Hence, by controlling the uniform generalization risk, one improves the generalization risk of $\mathcal{L}$ both in expectation and with a high probability.

The same proof technique used in Theorem 4 also implies the following concentration bound, which is useful when $I(\mathbf{h}; \mathbf{s}) = o(m)$ where $I(\mathbf{x}; \mathbf{y})$ is the Shannon mutual information. The following bound is similar to the bound derived by [23] using properties of sub-Gaussian loss functions.

Proposition 3. *Let $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ be a learning algorithm, whose risk is evaluated using a parametric loss function $l : \mathcal{Z} \times \mathcal{H} \to [0, 1]$. Then:*

$$p\Big\{\big|R_{\boldsymbol{s}}(\boldsymbol{h}) - R(\boldsymbol{h})\big| \geq t\Big\} \leq \frac{1}{t}\sqrt{\frac{I(\boldsymbol{s};\boldsymbol{h}) + 2}{2m}}.$$

Proof. The proof is in Appendix E. □

Note that having a vanishing mutual information, i.e., $I(\mathbf{s};\mathbf{h}) = o(m)$, which is the setting recently considered in the work of [23], is a *strictly stronger* condition than uniform generalization. For instance, we will later construct *deterministic* learning algorithms that generalize uniformly in expectation even though $I(\mathbf{s};\mathbf{h})$ is unbounded (see Theorem 8). By contrast, $I(\mathbf{s};\mathbf{h}) = o(m)$ is sufficient for $\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) \to 0$ to hold.

Finally, we note that the concentration bound depends linearly on the variational information $\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h})$. Typically, $\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) = O(1/\sqrt{m})$. By contrast, the VC bound provides an exponential decay on m [3,17]. Can the concentration bound in Theorem 4 be improved? The following proposition answers this question in the negative.

Proposition 4. *For any rational $0 < t < 1$, there exists a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$, a distribution $p(z)$, and a parametric loss $l : \mathcal{Z} \times \mathcal{H} \to [0, 1]$ such that:*

$$p\Big\{\big|R_{\boldsymbol{s}}(\boldsymbol{h}) - R(\boldsymbol{h})\big| = t\Big\} = \frac{\mathcal{J}(\hat{\boldsymbol{z}};\, \boldsymbol{h})}{t},$$

where the probability is evaluated over the random choice of $\boldsymbol{s}$ and the internal randomness of $\mathcal{L}$.

Proof. The proof is in Appendix F. □

Proposition 4 shows that, without making any additional assumptions beyond that of uniform generalization, the concentration bound in Theorem 4 is tight up to constant factors. Essentially, the only difference between the upper and the lower bounds is a vanishing $O(1/\sqrt{m})$ term that is *independent* of $\mathcal{L}$.

5. Properties of the Learning Capacity

In this section, we derive bounds on the learning capacity under various settings. We also describe some of its important properties.

5.1. Data Processing

The relationship between learning capacity and data processing is presented in Lemma 1. Given the random variables $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ and the Markov chain $\mathbf{x} \to \mathbf{y} \to \mathbf{z}$, we always have $\mathcal{J}(\mathbf{x};\, \mathbf{z}) \leq \mathcal{J}(\mathbf{x};\, \mathbf{y})$. Hence, we have a *partial order* on learning algorithms. This presents us with an important qualitative insight into the design of machine learning algorithms.

Suppose we have two different hypotheses $\mathbf{h}_1$ and $\mathbf{h}_2$. We will say that $\mathbf{h}_2$ contains *less information* than $\mathbf{h}_1$ if the Markov chain $\mathbf{s} \to \mathbf{h}_1 \to \mathbf{h}_2$ holds. For example, if the observations $\mathbf{z}_i \in \{0, 1\}$ are Bernoulli trials, then $\mathbf{h}_1 \in \mathbb{R}$ can be the empirical average as given in Example 1 while $\mathbf{h}_2 \in \{0, 1\}$ can be the label that occurs most often in the training set. Because $\mathbf{h}_2 = \mathbb{I}\{\mathbf{h}_1 \geq m/2\}$, the hypothesis $\mathbf{h}_2$ contains strictly less information about the original training set than $\mathbf{h}_1$. Formally, we have $\mathbf{s} \to \mathbf{h}_1 \to \mathbf{h}_2$. In this case, $\mathbf{h}_2$ enjoys a better *uniform* generalization bound because of data-processing. Intuitively, we know that such a result should hold because $\mathbf{h}_2$ is less dependent to the original training set than $\mathbf{h}_1$. Hence, one can improve the uniform generalization bound (or equivalently the learning capacity) of a learning algorithm

by post-processing its hypothesis $\mathbf{h}$ in a manner that is conditionally independent of the original training set given $\mathbf{h}$.

Example 6. *Post-processing hypotheses is a common technique in machine learning. This includes sparsifying the coefficient vector $\boldsymbol{w} \in \mathbb{R}^d$ in linear methods, where $\boldsymbol{w}_j$ is set to zero if it has a small absolute magnitude. It also includes methods that have been proposed to reduce the number of support vectors in SVM by exploiting linear dependence [30], or some methods for decision tree pruning. By the data processing inequality, such techniques reduce the learning capacity and, as a consequence, mitigate the risk for overfitting.*

Needless to mention, better generalization does not immediately translate into a smaller true risk. This is because the empirical risk itself may increase when the hypothesis $\mathbf{h}$ is post-processed *independently* of the original training sample.

5.2. Effective Domain Size

Next, we look into how the size of the domain $\mathcal{Z}$ limits the learning capacity. First, we start with the following definition:

Definition 7 (Lazy Learning)**.** *A learning algorithm $\mathcal{L}$ is called* lazy *if the training sample $\boldsymbol{s} \in \mathcal{Z}^m$ can be reconstructed perfectly from the hypothesis $\boldsymbol{h} \in \mathcal{H}$. In other words, $H(\boldsymbol{s}|\boldsymbol{h}) = 0$, where H is the Shannon entropy. Equivalently, the mapping from $\boldsymbol{s}$ to $\boldsymbol{h}$ is injective.*

One common example of a lazy learner is instance-based learning when $\mathbf{h} = \mathbf{s}$. Despite their simple nature, lazy learners are useful in practice. They are useful theoretical tools as well. In particular, because of the fact that $H(\mathbf{s}|\mathbf{h}) = 0$ and the data processing inequality, the learning capacity of a lazy learner provides an upper bound to the learning capacity of *any* possible learning algorithm. Therefore, we can relate the learning capacity $C(\mathcal{L})$ to the size of the domain $\mathcal{Z}$ by determining the learning capacity of lazy learners. Because the size of $\mathcal{Z}$ is usually infinite, we introduce the following definition of *effective* set size.

Definition 8. *In a countable space $\mathcal{Z}$ endowed with a probability mass function $p(z)$, the effective size of $\mathcal{Z}$ w.r.t. $p(z)$ is defined by:* $\mathbf{Ess}_{p(z)}\left(\mathcal{Z}\right) \doteq 1 + \left(\sum_{z \in \mathcal{Z}} \sqrt{p(z)\left(1 - p(z)\right)}\right)^2$.

At one extreme, if $p(z)$ is *uniform* over a finite alphabet $\mathcal{Z}$, then $\mathbf{Ess}_{p(z)}\left(\mathcal{Z}\right) = |\mathcal{Z}|$. At the other extreme, if $p(z)$ is a Kronecker delta distribution, then $\mathbf{Ess}_{p(z)}\left(\mathcal{Z}\right) = 1$. As proved next, this notion of effective set size *determines* the rate of convergence of an empirical probability mass function to its true distribution when the distance is measured in the total variation sense. As a result, it allows us to relate the learning capacity to a property of the domain $\mathcal{Z}$.

Theorem 5. *Let $\mathcal{Z}$ be a countable space endowed with a probability mass function $p(z)$. Let $\boldsymbol{s}$ be a set of m i.i.d. observations $\mathbf{z}_i \sim p(z)$. Define $p_{\boldsymbol{s}}(z)$ to be the empirical probability mass function that results from drawing observations uniformly at random from $\boldsymbol{s}$. Then:*

$$\mathbb{E}_{\boldsymbol{s}}\, ||p(z),\, p_{\boldsymbol{s}}(z)||_{\mathcal{T}} = \sqrt{\frac{\mathbf{Ess}_{p(z)}\left[\mathcal{Z}\right] - 1}{2\,\pi\,m}} + o(1/\sqrt{m}),$$

where $\mathbf{Ess}_{p(z)}\left[\mathcal{Z}\right]$ is the effective size of $\mathcal{Z}$ (see Definition 8).

Proof. The proof is in Appendix G. □

A special case of Theorem 5 was proved by de Moivre in the 1730s, who showed that the empirical mean of i.i.d. Bernoulli trials with a probability of success ϕ converges to the true mean with rate $\sqrt{2\phi(1-\phi)/(\pi m)}$. This is believed to be the first appearance of the square-root law in statistical inference in the literature [31]. Because the effective domain size of the Bernoulli distribution, according to Definition 8, is given by $1 + 4\phi(1-\phi)$, Theorem 5 agrees with, in fact generalizes, de Moivre's result.

Corollary 1. *Let $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ be a learning algorithm whose hypothesis is $\boldsymbol{h} \in \mathcal{H}$. Then, $\mathcal{J}(\hat{\mathbf{z}}; \boldsymbol{h}) \leq \sqrt{\frac{\mathbf{Ess}_{p(z)}[\mathcal{Z}]-1}{2\,\pi\,m}} + o(1/\sqrt{m})$. Moreover, the bound is achieved by lazy learners.*

Proof. Let $\tilde{\mathbf{h}}$ be the hypothesis produced by a lazy learner. The simplest example is if $\mathbf{h}$ is equal to the training sample $\mathbf{s}$ itself. Then, we always have the Markov chain $\mathbf{s} \to \tilde{\mathbf{h}} \to \mathbf{h}$ for any hypothesis $\mathbf{h} \in \mathcal{H}$. Therefore, by the data processing inequality, we have $\mathcal{J}(\hat{\mathbf{z}}; \mathbf{h}) \leq \mathcal{J}(\hat{\mathbf{z}}; \tilde{\mathbf{h}})$. By Theorem 5, we have:

$$\mathcal{J}(\hat{\mathbf{z}}; \tilde{\mathbf{h}}) = \sqrt{\frac{\mathbf{Ess}_{p(z)}[\mathcal{Z}]-1}{2\,\pi\,m}} + o(1/\sqrt{m})$$

Hence, the statement of the corollary follows. □

Corollary 2. *For any learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$, we have $C(\mathcal{L}) \leq \sqrt{\frac{|\mathcal{Z}|-1}{2\,\pi\,m}} + o(1/\sqrt{m})$.*

Proof. The function $f(p) = \sum_z \sqrt{p(z)(1-p(z))}$ is both concave over the probability simplex and permutation-invariant. Hence, by symmetry, the maximum effective domain size must be achieved at the uniform distribution $p(z) = 1/|\mathcal{Z}|$, in which case $\mathbf{Ess}_{p(z)}[\mathcal{Z}] = |\mathcal{Z}|$. □

5.3. Finite Hypothesis Space

Next, we look into the role of the *size* of the hypothesis space. This is formalized by the following theorem.

Theorem 6. *Let $\boldsymbol{h} \in \mathcal{H}$ be the hypothesis produced by a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$. Then:*

$$C(\mathcal{L}) \leq \sqrt{\frac{\mathrm{H}(\boldsymbol{h})}{2\,m}} \leq \sqrt{\frac{\log|\mathcal{H}|}{2\,m}},$$

where H *is the Shannon entropy measured in nats.*

Proof. If we let $I(\mathbf{x}; \mathbf{y})$ be the mutual information between the r.v.'s $\mathbf{x}$ and $\mathbf{y}$ and let $\mathbf{s} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}$ be the training set, we have:

$$\begin{aligned} I(\mathbf{s}; \mathbf{h}) &= H(\mathbf{s}) - H(\mathbf{s}\,|\,\mathbf{h}) \\ &= \Big[\sum_{i=1}^{m} H(\mathbf{z}_i)\Big] - \Big[H(\mathbf{z}_1|\mathbf{h}) + H(\mathbf{z}_2|\mathbf{z}_1, \mathbf{h}) + \cdots\Big] \end{aligned}$$

Because conditioning reduces entropy, i.e., $H(\mathbf{x}|\mathbf{y}) \leq H(\mathbf{x})$ for any r.v.'s $\mathbf{x}$ and $\mathbf{y}$, we have:

$$I(\mathbf{s}; \mathbf{h}) \geq \sum_{i=1}^{m} [H(\mathbf{z}_i) - H(\mathbf{z}_i\,|\,\mathbf{h})] = m\,[H(\hat{\mathbf{z}}) - H(\hat{\mathbf{z}}\,|\,\mathbf{h})]$$

Therefore:

$$I(\hat{\mathbf{z}};\, \mathbf{h}) \leq \frac{I(\mathbf{s};\, \mathbf{h})}{m} \tag{13}$$

Next, we use *Pinsker's inequality* [10], which states that for any probability measures p and q: $||p\,,\ q||_{\mathcal{T}} \leq \sqrt{\frac{D(p\,||\,q)}{2}}$, where $||p\,,\ q||_{\mathcal{T}}$ is total variation distance and $D(p\,||\,q)$ is the Kullback-Leibler divergence measured in nats. If we recall that $\mathcal{J}(\mathbf{s};\, \mathbf{h}) = ||p(\mathbf{s})\, p(\mathbf{h})\,,\ p(\mathbf{s},\, \mathbf{h})||_{\mathcal{T}}$ while the mutual information is $I(\mathbf{s};\mathbf{h}) = D(p(\mathbf{s},\mathbf{h})\,||\,p(\mathbf{s})\,p(\mathbf{h}))$, we deduce from Pinsker's inequality and Equation (13):

$$\begin{aligned}\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) &= ||p(\hat{\mathbf{z}})\, p(\mathbf{h})\,,\ p(\hat{\mathbf{z}},\mathbf{h})||_{\mathcal{T}} \\ &\leq \sqrt{\frac{I(\hat{\mathbf{z}};\mathbf{h})}{2}} \leq \sqrt{\frac{I(\mathbf{s};\mathbf{h})}{2m}} \leq \sqrt{\frac{H(\mathbf{h})}{2m}} \leq \sqrt{\frac{\log|\mathcal{H}|}{2m}}. \quad \square\end{aligned}$$

Theorem 6 re-establishes the classical PAC result on the finite hypothesis space setting. However, unlike its typical proofs, the proof presented here is purely information-theoretic and does not make any references to the union bounds.

5.4. Differential Privacy

Randomization reduces the risk for overfitting. One common randomization technique in machine learning is differential privacy [32,33], which addresses the goal of obtaining useful information about the sample $\mathbf{s}$ as a whole without revealing a lot of information about any individual observation. Here, we show that differentially-private learning algorithms have small learning capacities.

Definition 9 ([33]). *A randomized learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ is (ϵ, δ) differentially private if for any $\mathcal{O} \subseteq \mathcal{H}$ and any two samples $\boldsymbol{s}$ and $\boldsymbol{s}'$ that differ in one observation only, we have:*

$$p(\boldsymbol{h} \in \mathcal{O} \mid \boldsymbol{s}) \leq e^{\epsilon} \cdot p(\boldsymbol{h} \in \mathcal{O} \mid \boldsymbol{s}') + \delta$$

Proposition 5. *If a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ is (ϵ, δ) differentially private, then: $\mathcal{J}(\hat{\boldsymbol{z}};\, \boldsymbol{h}) \leq (e^{\epsilon} - 1 + \delta)/2$.*

Proof. The proof is in Appendix H. $\square$

Not surprisingly, the differential privacy parameters (ϵ, δ) control the uniform generalization risk, where small values of ϵ and δ lead to a reduced risk for overfitting.

5.5. Empirical Risk Minimization of 0–1 Loss Classes

Empirical risk minimization (ERM) of stochastic loss is a popular approach for learning from data. It is often regarded as the default strategy to use, due to its simplicity, generality, and statistical efficiency [1,3,13,34]. Given a fixed hypothesis space $\mathcal{H}$, a domain $\mathcal{Z}$, and a loss function $l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, the ERM learning rule selects the hypothesis $\hat{\mathbf{h}}_{\mathbf{s}}$ that minimizes the empirical risk:

$$\hat{\mathbf{h}}_{\mathbf{s}} = \arg\min_{h \in \mathcal{H}} \left\{ L_{\mathbf{s}}(h) = \frac{1}{|\mathbf{s}|} \sum_{\mathbf{z}_i \in \mathbf{s}} l(\mathbf{z}_i, h) \right\}, \tag{14}$$

By contrast, the true risk minimizer $\mathbf{h}^{\star}$ is:

$$\mathbf{h}^{\star} = \arg\min_{h \in \mathcal{H}} \left\{ L(h) = \mathbb{E}_{\mathbf{z} \sim p(z)} \left[l(\mathbf{z},\, h) \right] \right\}. \tag{15}$$

Hence, learning via ERM is justified if $L(\hat{\mathbf{h}}_{\mathbf{s}}) \leq L(\mathbf{h}^{\star}) + \epsilon$, for some $\epsilon \ll 1$. If such a condition holds and $\epsilon \to 0$ as the sample size m increases, the ERM learning rule is called *consistent*.

Uniform generalization is a sufficient condition for the consistency of empirical risk minimization (ERM). To see this, we have by definition:

$$\begin{aligned}\mathbb{E}_{\mathbf{s}}[L_{\mathbf{s}}(\hat{\mathbf{h}}_{\mathbf{s}})] &= \mathbb{E}_{\mathbf{s}}[\min_{h\in\mathcal{H}} L_{\mathbf{s}}(h)] \\ &\leq \min_{h\in\mathcal{H}} \big\{\mathbb{E}_{\mathbf{s}}[L_{\mathbf{s}}(h)]\big\} = \min_{h\in\mathcal{H}} L(h) = R(\mathbf{h}^{\star}),\end{aligned}$$

From this, we conclude that:

$$\mathbb{E}_{\mathbf{s}} R(\hat{\mathbf{h}}_{\mathbf{s}}) - R(\mathbf{h}^{\star}) \leq \mathbb{E}_{\mathbf{s}} R(\hat{\mathbf{h}}_{\mathbf{s}}) - \mathbb{E}_{\mathbf{s}}[L_{\mathbf{s}}(\hat{\mathbf{h}}_{\mathbf{s}})] \leq C(\mathcal{L}),$$

where $C(\mathcal{L})$ is the learning capacity of the empirical risk minimization rule. The last inequality follows from Theorem 2. In addition, because $R(\hat{\mathbf{h}}_{\mathbf{s}}) - R(\mathbf{h}^{\star}) \geq 0$, we have by the Markov inequality:

$$p_{\mathbf{s}}\Big\{R(\hat{\mathbf{h}}_{\mathbf{s}}) - R(\mathbf{h}^{\star}) \geq t\Big\} \leq \frac{\mathbb{E}_{\mathbf{s}} R(\hat{\mathbf{h}}_{\mathbf{s}}) - R(\mathbf{h}^{\star})}{t} \leq \frac{C(\mathcal{L})}{t}$$

Hence, the ERM learning rule is consistent if $C(\mathcal{L}) \to 0$ as $m \to \infty$. Next, we describe when such a condition on $C(\mathcal{L})$ holds for 0–1 loss classes. To do that, we begin with two familiar definitions from statistical learning theory.

Definition 10 (Shattered Set)**.** *Given a domain $\mathcal{Z}$, a hypothesis space $\mathcal{H}$, and a 0–1 loss function $l : \mathcal{Z} \times \mathcal{H} \to \{0,1\}$, a set $\{z_1, \ldots, z_d\}$ is said to be shattered by $\mathcal{H}$ with respect to the function l if for any labeling $I \in \{0,1\}^d$, there exists a hypothesis $h_I \in \mathcal{H}$ such that $(l(z_1, h_I), \ldots, l(z_d, h_I)) = I$.*

Example 7. *Let $\mathcal{Z} = \mathcal{H} = \mathbb{R}$ and let the loss function be $l(z,h) = \mathbb{I}\{z - h \geq 0\}$. Then, any singleton set $\{z\}$ is shattered by $\mathcal{H}$ since we always have the two hypotheses $h_0 = z - 1$ and $h_1 = z + 1$. However, no set of two points in $\mathcal{Z}$ can be shattered by $\mathcal{H}$. By contrast, if the hypothesis is a pair $(h, c) \in \mathbb{R} \times \mathbb{R}$ and the loss function is $l(z, h, b) = \mathbb{I}\{c\, z - h \geq 0\}$, then any set of two distinct examples $\{z_1, z_2\}$ is shattered by the hypothesis space.*

Definition 11 (VC Dimension)**.** *The VC dimension of a hypothesis space $\mathcal{H}$ with respect to a domain $\mathcal{Z}$ and a 0–1 loss $l : \mathcal{Z} \times \mathcal{H} \to \{0,1\}$ is the maximum cardinality of a set of points in $\mathcal{Z}$ that can be shattered by $\mathcal{H}$ with respect to l.*

The VC dimension is arguably the most fundamental concept in statistical learning theory because it provides a crisp characterization of learnability for 0–1 loss classes. Next, we show that the VC dimension has, in fact, an equivalence characterization with the learning capacity $C(\mathcal{L})$. Specifically, under the Axiom of Choice, an ERM learning rule exists that has a vanishing learning capacity $C(\mathcal{L})$ if and only if the 0–1 loss class has a finite VC dimension.

Before we establish this important result, we describe why ERM by itself is not sufficient for uniform generalization to hold even when the hypothesis space has a finite VC dimension.

Proposition 6. *For any sample size $m \geq 1$ and a positive constant $\epsilon > 0$, there exists a hypothesis space $\mathcal{H}$, a domain $\mathcal{Z}$, and a 0–1 loss $l : \mathcal{Z} \times \mathcal{H} \to \{0,1\}$ such that: (1) $\mathcal{H}$ has a VC dimension $d = 1$, and (2) a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ exists that outputs an empirical risk minimizer $\hat{\boldsymbol{h}}_{\boldsymbol{s}}$ with $\mathcal{J}(\hat{\mathbf{z}}; \hat{\boldsymbol{h}}_{\boldsymbol{s}}) \geq 1 - \epsilon$.*

Proof. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = \{+1, -1\}$ and let the loss be $l(x, y, h) = \mathbb{I}\{y \cdot (x - h) \leq 0\}$. In other words, the goal is to learn a threshold in the unit interval that separates the positive from the negative examples. Let $\mathbf{x} \in \mathcal{X}$ be uniformly distributed in $[0, 1]$ and let $\mathbf{h}^\star$ be an error-free separator. Then, for any training sample $\mathbf{s} \in \mathcal{Z}^m$, the set of all empirical risk minimizers $\hat{\mathbf{H}}$ is:

$$\hat{\mathbf{H}} = \big\{h \in [0,1] : \, y_i = \mathbf{sign}(x_i - h), \quad \forall i \in \{1, \ldots, m\}\big\}$$

In particular, $\hat{\mathbf{H}}$ is an interval, which has the power of the continuum, so it can be used to encode the entire training sample.

Fix $\delta > 0$ in advance, which can be made arbitrarily small. Then, the probability over the random choice of the sample that $|\hat{\mathbf{H}}| < \delta$ can be made arbitrarily small for a sufficiently small $\delta > 0$, where $|\hat{\mathbf{H}}|$ is the length of the interval.

Let $\hat{\mathbf{h}} \in \hat{\mathbf{H}}$ be a hypothesis that lies at the middle of $\hat{\mathbf{H}}$, i.e.,:

$$\hat{\mathbf{h}} = \frac{1}{2}\Big[\arg \max_{x_i \in \mathbf{s} \wedge y_i = -1} x_i + \arg \min_{x_i \in \mathbf{s} \wedge y_i = +1} x_i\Big]$$

Let $k = 1 + \log_2(1/\delta)$. Then, $[\hat{\mathbf{h}} - 2^{-k}, \hat{\mathbf{h}} + 2^{-k}] \subseteq \hat{\mathbf{H}}$ holds with a high probability (which can be made arbitrarily close to 1 for a sufficiently small δ). Let $\tilde{\mathbf{h}}$ be a hypothesis whose binary expansion agrees with $\hat{\mathbf{h}}$ in its first $k + 1$ bits and encodes the entire training sample in the rest of the bits.

Finally, the output of the learning algorithm is $\hat{\mathbf{h}}_\mathbf{s}$, which is given by the following rule:

1. If $\tilde{\mathbf{h}}$ is an empirical risk minimizer, then set $\hat{\mathbf{h}}_\mathbf{s} = \tilde{\mathbf{h}}$
2. Otherwise, set $\hat{\mathbf{h}}_\mathbf{s} = \hat{\mathbf{h}}$.

Now, define the following *different* parametric loss $l' : \mathcal{Z} \to [0, 1]$ to be a function that first uses $\hat{\mathbf{h}}_\mathbf{s}$ to *decode* the training sample $\mathbf{s}$ based on the coding method constructed above and, then, assigns 1 if and only if $x \in \mathbf{s}$. To reiterate, this decoding succeeds with a probability that can be made arbitrarily high for a sufficiently small $\delta > 0$. Clearly, l' is a loss defined on the product space $\mathcal{Z} \times \mathcal{H}$ and has a bounded range. However, the generalization risk w.r.t. l' is, at least, equal to the probability that $|\hat{\mathbf{H}}| < \delta$, which can be made arbitrarily close to 1. Hence, the statement of the proposition holds. □

Proposition 6 shows that one cannot obtain a non-trivial bound on the uniform generalization risk of an ERM learning rule in terms of the VC dimension d and the sample size m without making some additional assumptions. Next, we prove that an ERM learning rule *exists* that satisfies the uniform generalization property if the hypothesis space has a finite VC dimension. We begin by recalling a fundamental result in modern set theory. A non-empty set $\mathcal{Q}$ is said to be *well-ordered* if $\mathcal{Q}$ is endowed with a total order $\preceq$ such that every non-empty subset of $\mathcal{Q}$ contains a least element. The following fundamental result, which was published in 1904, is due to Ernst Zermelo [35].

Theorem 7 (Well-Ordering Theorem)**.** *Under the Axiom of Choice, every non-empty subset can be well-ordered.*

Theorem 8. *Given a hypothesis space $\mathcal{H}$, a domain $\mathcal{Z}$, and a 0–1 loss $l : \mathcal{H} \times \mathcal{Z} \to \{0, 1\}$, let $\preceq$ be a well-ordering on $\mathcal{H}$ and let $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ be the learning rule that outputs the " least" empirical risk minimizer to the training sample $\mathbf{s} \in \mathcal{Z}^m$ according to $\preceq$. Then, $C(\mathcal{L}) \to 0$ as $m \to \infty$ if $\mathcal{H}$ has a finite VC dimension. In particular:*

$$C(\mathcal{L}) \leq \frac{3}{\sqrt{m}} + \sqrt{\frac{1 + d \log \frac{2em}{d}}{m}},$$

where d is the VC dimension of $\mathcal{H}$, provided that $m \geq d$.

Proof. The proof is in Appendix I. □

Next, we prove a converse statement. Before we do this, we present a learning problem that shows why a converse to Theorem 8 is not generally possible without making some additional assumptions. Hence, our converse will be later established for the binary classification setting only.

Example 8 (Subset Learning Problem)**.** *Let $\mathcal{Z} = \{1, 2, 3, \ldots, d\}$ be a finite set of positive integers. Let $\mathcal{H} = 2^{\mathcal{Z}}$ and define the 0–1 loss of a hypothesis $h \in \mathcal{H}$ to be $l(z, h) = \mathbb{I}\{z \notin h\}$. Then, the VC dimension is d. However, the learning rule that outputs $h = \mathcal{Z}$ is always an ERM learning rule that generalizes uniformly with rate $\epsilon = 0$ regardless of the sample size and the distribution of observations.*

The previous example shows that a converse to Theorem 8 is not generally possible without making some additional assumptions. In particular, in the Subset Learning Problem, the VC dimension is not an accurate measure of the complexity of the hypothesis space $\mathcal{H}$ because many hypotheses dominate others (i.e., perform better across all distributions of observations). For example, the hypothesis $h' = \{1, 2, 3\}$ dominates $h'' = \{1\}$ because there is no distribution on observations in which h'' outperforms h'. In fact, the hypothesis $h = \mathcal{Z}$ dominates all other hypotheses.

Consequently, in order to prove a lower bound for all ERM rules, we focus on the standard binary classification setting.

Theorem 9. *In any fixed domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, let the hypothesis space $\mathcal{H}$ be a concept class on $\mathcal{X}$ and let $l(x, y, h) = \mathbb{I}\{y \neq h(x)\}$ be the misclassification error. Then, any ERM learning rule $\mathcal{L}$ w.r.t. l has a learning capacity $C(\mathcal{L})$ that is bounded from below by $C(\mathcal{L}) \geq \frac{1}{2}\left(1 - \frac{1}{d}\right)^m$, where m is the training sample size and d is the VC dimension of $\mathcal{H}$.*

Proof. The proof is in Appendix J. □

Using both Theorems 8 and 9, we arrive at the following equivalence characterization of the VC dimension of a concept class with the learning capacity.

Theorem 10. *Given a fixed domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, let the hypothesis space $\mathcal{H}$ be a concept class on $\mathcal{X}$ and let $l(x, y, h) = \mathbb{I}\{y \neq h(x)\}$ be the misclassification error. Let m be the sample size. Then, the following statements are equivalent under the Axiom of Choice:*

1. *$\mathcal{H}$ admits an ERM learning rule $\mathcal{L}$ whose learning capacity $C(\mathcal{L})$ satisfies $C(\mathcal{L}) \to 0$ as $m \to \infty$.*
2. *$\mathcal{H}$ has a finite VC dimension.*

Proof. The lower bound in Theorem 9 holds for all ERM learning rules. Hence, an ERM learning rule exists that generalize uniformly with a vanishing rate across all distributions only if $\mathcal{H}$ has a finite VC dimension. However, under the Axiom of Choice, $\mathcal{H}$ can always be well-ordered by Theorem 7 so, by Theorem 8, a finite VC dimension is also sufficient to guarantee the existence of a learning rule that generalize uniformly. □

Theorem 10 presents a characterization of the VC dimension in terms of information theory. According to the theorem, an ERM learning rule can be constructed that does not encode the training sample *if and only if* the hypothesis space has a finite VC dimension.

Remark 2. *One method of constructing a well-ordering on a hypothesis space $\mathcal{H}$ is to use the fact that computers are equipped with finite precisions. Hence, in practice, every hypothesis space is enumerable, from which the normal ordering of the integers forms a valid well-ordering on $\mathcal{H}$.*

6. Concluding Remarks

In this paper, we introduced the notion of " learning capacity" for algorithms that learn from data, which is analogous to the Shannon capacity of communication channels. Learning capacity is an information-theoretic quantity that measures the contribution of a single training example to the final hypothesis. It has three equivalent interpretations: (1) as a tight upper bound on the uniform generalization risk, (2) as a measure of information leakage, and (3) as a measure of algorithmic stability. Furthermore, by establishing a chain rule for learning capacity, concentration bounds were derived, which revealed that the learning capacity controlled both the expectation of the generalization risk and its variance. Moreover, the relationship between algorithmic stability and data processing revealed that algorithmic stability can be improved by post-processing the learned hypothesis.

Throughout this paper, we provided several bounds on the learning capacity under various settings. For instance, we established a relationship between algorithmic stability and the effective size of the domain of observations, which can be interpreted as a formal justification for dimensionality reduction methods. Moreover, we showed how learning capacity recovered classical bounds, such as in the finite hypothesis space setting, and derived new bounds for other settings as well, such as differential privacy. We also established that, under the Axiom of Choice, the existence of an empirical risk minimization (ERM) rule for 0–1 loss classes that had a vanishing learning capacity was equivalent to the assertion that the hypothesis space had a finite Vapnik–Chervonenkis (VC) dimension, thus establishing an equivalence relation between two of the most fundamental concepts in statistical learning theory and information theory.

More generally, the intent of this work is to bring to light a new information-theoretic approach for analyzing machine learning algorithms. Despite the fact that " uniform generalization" might appear to be a strong condition at a first sight, one of the central claims of this paper is that uniform generalization is, in fact, a natural condition that arises commonly in practice. It is not a condition to require or enforce! We believe this holds because any learning algorithm is a *channel* from the space of training samples to the hypothesis space. Because learning is a mapping between two spaces, its risk for overfitting should be determined from the mapping itself (i.e., independently of the choice of the loss function). Such an approach yields the uniform generalization bounds that are derived in this paper.

It is worth highlighting that uniform generalization bounds can be established for many other settings that have not be discussed in this paper and it has found some promising applications. Using sample compression schemes, one can show that any learnable hypothesis space is also learnable by an algorithm that achieves uniform generalization [36]. Also, generalization bounds for stochastic convex optimization yield information criteria for model selection that can outperform the popular Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC) [37]. More recently, uniform generalization has inspired the development of new approaches for structured regression as well [38].

7. Further Research Directions

Before we conclude, we suggest future directions of research and list some open problems.

7.1. Induced VC Dimension

The variational information $\mathcal{J}(\hat{\mathbf{z}}; \mathbf{h})$ provides an upper bound on the generalization risk of the learning algorithm $\mathcal{L}$ across all parametric loss classes. This upper bound is *achievable* by the generalization risk of the *binary reconstruction loss*:

$$l(z, \mathbf{h}) = \mathbb{I}\{p(z \in \mathbf{s} \mid \mathbf{h}) \ \geq \ p(z \in \mathbf{s})\}, \tag{16}$$

which assigns the value one to observations $z \in \mathcal{Z}$ that are *more* likely to have been present in the training sample $\mathbf{s}$ upon knowing $\mathbf{h}$, and assigns zero otherwise. In expectation, the generalization risk of this parametric loss is the worst generalization risk across all parametric loss classes.

Let both $p(z)$ and $p(h|z)$ be fixed; the first is the distribution of observations while the second is entirely determined by the learning algorithm $\mathcal{L}$. Then, because the loss in Equation (16) is binary, it has a VC dimension, which we will call the *induced VC dimension* of the learning algorithm $\mathcal{L}$ [39]. Note that this induced VC dimension is defined for all learning problems, including regression and clustering, but it is *distribution-dependent*, which is quite unlike the traditional VC dimension of hypothesis spaces.

There are a lot of open questions related to the *induced VC dimension* of learning algorithms. For instance, while a finite VC dimension implies a small variational information, when does the converse also hold? Can we obtain a non-trivial bound on the induced VC dimension of a learning algorithm $\mathcal{L}$ upon knowing its uniform generalization risk $\mathcal{J}(\hat{\mathbf{z}}; \mathbf{h})$? Along similar lines, suppose that $\mathcal{L}$ is an empirical risk minimization (ERM) algorithm of a 0–1 loss class that may or may not use an appropriate tie breaking rule (in light of what was discussed in Section 5.5). Is there a non-trivial relation between the VC dimension of the 0–1 loss that is being minimized and the induced VC dimension of the ERM learning algorithm?

7.2. Unsupervised Model Selection

Information criteria (such as AIC and BIC), are sometimes used in the unsupervised learning setting for model selection, such as when determining the value of k in the popular k-means algorithm [40]. Given that the notion of uniform generalization is developed in the *general* setting of learning, should the learning capacity $C(\mathcal{L})$ serve as a model selection criterion in the unsupervised setting? Why or why not?

7.3. Effective Domain Size

The effective size of the domain of a random variable $\mathbf{z}$ in Definition 8 satisfies some intuitive properties and violates others. For instance, it reduces to the size of the domain $|\mathcal{Z}|$ when the distribution is uniform. Moreover, if $\mathbf{z}$ is Bernoulli, the effective domain size is determined by the *variance* of the Bernoulli distribution. Importantly, this notion is well-motivated because it determines the rate of convergence of an empirical probability mass function to its true distribution when the distance is measured in the total variation sense. As a result, it allowed us to relate the learning capacity to a property of the domain $\mathcal{Z}$.

However, such a notion of effective domain size has some surprising properties. For instance, the effective size of the domain of two *independent* random variables is not equal to the product of the effective size of each individual domain! In rate distortion theory, a similar phenomenon is observed. Reference [10] explain this observation by stating that " rectangular grid points (arising from independent descriptions) do not fill up the space efficiently." Can the effective domain size in Definition 8 be motivated using rate distortion theory?

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Lemma 2

With no loss of generality, let's assume that all domains are enumerable. We have:

$$\begin{aligned}\mathcal{J}(\mathbf{x};\,(\mathbf{y},\mathbf{z})) &= 1-\sum_{x,y,z}\min\left\{p(\mathbf{x}=x)\,p(\mathbf{y}=y,\mathbf{z}=z),\ p(\mathbf{x}=x,\mathbf{y}=y,\mathbf{z}=z)\right\}\\ &= 1-\sum_{x}p(\mathbf{x}=x)\sum_{y,z}\min\left\{p(\mathbf{y}=y,\mathbf{z}=z),\ p(\mathbf{y}=y,\mathbf{z}=z|\mathbf{x}=x)\right\}\end{aligned}$$

However, the minimum of the sums is always larger than the sum of minimums. That is:

$$\min\left\{\sum_i \alpha_i,\ \sum_i \beta_i\right\} \geq \sum_i \min\{\alpha_i,\ \beta_i\}$$

Using marginalization $p(\mathbf{x}) = \sum_y p(\mathbf{x},\,\mathbf{y}=y)$ and the above inequality, we obtain:

$$\begin{aligned}\mathcal{J}(\mathbf{x};\,(\mathbf{y},\mathbf{z})) &\geq 1-\sum_{x}p(\mathbf{x}=x)\sum_{y}\min\{\sum_{z}p(\mathbf{y}=y,\mathbf{z}=z),\ \sum_{z}p(\mathbf{y}=y,\mathbf{z}=z|\mathbf{x}=x)\}\\ &= 1-\sum_{x}p(\mathbf{x}=x)\sum_{y}\min\{p(\mathbf{y}=y),\ p(\mathbf{y}=y|\mathbf{x}=x)\}\\ &= \mathcal{J}(\mathbf{x};\,\mathbf{y})\end{aligned}$$

Appendix B. Proof of Theorem 1

We will first prove the inequality when $k=2$. First, we write by definition:

$$\mathcal{J}(\mathbf{z};\,(\mathbf{h}_1,\mathbf{h}_2)) = ||p(\mathbf{z},\mathbf{h}_1,\mathbf{h}_2)\,,\ \ p(\mathbf{z})\,p(\mathbf{h}_1,\mathbf{h}_2)||_{\mathcal{T}}$$

Using the fact that the total variation distance is related to the ℓ_1 distance by $||P,\ \ Q||_{\mathcal{T}} = \frac{1}{2}||P-Q||_1$, we have:

$$\begin{aligned}\mathcal{J}(\mathbf{z};\,(\mathbf{h}_1,\mathbf{h}_2)) &= \frac{1}{2}\big|\big|\ p(\mathbf{z},\mathbf{h}_1,\mathbf{h}_2) - p(\mathbf{z})\,p(\mathbf{h}_1,\mathbf{h}_2)\ \big|\big|_1\\ &= \frac{1}{2}\big|\big|\ p(\mathbf{z},\mathbf{h}_1)\,p(\mathbf{h}_2|\mathbf{z},\mathbf{h}_1) - p(\mathbf{z})\,p(\mathbf{h}_1)\,p(\mathbf{h}_2|\mathbf{h}_1)\ \big|\big|_1\\ &= \frac{1}{2}\big|\big|\ \left[p(\mathbf{z},\mathbf{h}_1) - p(\mathbf{z})\,p(\mathbf{h}_1)\right]\cdot p(\mathbf{h}_2|\mathbf{h}_1)\\ &\qquad + p(\mathbf{z},\mathbf{h}_1)\ \cdot\left[p(\mathbf{h}_2|\mathbf{z},\mathbf{h}_1) - p(\mathbf{h}_2|\mathbf{h}_1)\right]\ \big|\big|_1\end{aligned}$$

Using the triangle inequality:

$$\mathcal{J}(\mathbf{z};\,(\mathbf{h}_1,\mathbf{h}_2)) \leq \frac{1}{2}\Big|\Big|\left[p(\mathbf{z},\mathbf{h}_1) - p(\mathbf{z})\,p(\mathbf{h}_1)\right]\cdot p(\mathbf{h}_2|\mathbf{h}_1)\Big|\Big|_1 + \frac{1}{2}\Big|\Big|p(\mathbf{z},\mathbf{h}_1)\cdot\left[p(\mathbf{h}_2|\mathbf{z},\mathbf{h}_1) - p(\mathbf{h}_2|\mathbf{h}_1)\right]\Big|\Big|_1$$

The above inequality is interpreted by expanding the ℓ_1 distance into a sum of absolute values of terms in the product space $\mathcal{Z}\times\mathcal{H}_1\times\mathcal{H}_2$, where $\mathbf{h}_k\in\mathcal{H}_k$. Next, we bound each term on the right-hand side separately. For the first term, we note that:

$$\frac{1}{2}\big|\big|\ \left[p(\mathbf{z},\mathbf{h}_1) - p(\mathbf{z})\,p(\mathbf{h}_1)\right]\cdot p(\mathbf{h}_2|\mathbf{h}_1)\ \big|\big|_1 = \frac{1}{2}\big|\big|\ p(\mathbf{z},\mathbf{h}_1) - p(\mathbf{z})\,p(\mathbf{h}_1)\ \big|\big|_1 = \mathcal{J}(\mathbf{z};\,\mathbf{h}_1) \tag{A1}$$

The equality holds by expanding the ℓ_1 distance and using the fact that $\sum_{\mathbf{h}_2} p(\mathbf{h}_2|\mathbf{h}_1) = 1$.

However, the second term can be re-written as:

$$\begin{aligned}
&\frac{1}{2}\big|\big|\ p(\mathbf{z},\mathbf{h}_1)\cdot\left[p(\mathbf{h}_2|\mathbf{z},\mathbf{h}_1) - p(\mathbf{h}_2|\mathbf{h}_1)\right]\ \big|\big|_1 \\
&= \frac{1}{2}\big|\big|\ p(\mathbf{h}_1)\cdot\left[p(\mathbf{h}_2,\mathbf{z}|\mathbf{h}_1) - p(\mathbf{z}|\mathbf{h}_1)\,p(\mathbf{h}_2|\mathbf{h}_1)\right]\ \big|\big|_1 \\
&= \mathbb{E}_{\mathbf{h}_1}\left[||p(\mathbf{h}_2,\mathbf{z}|\mathbf{h}_1)\,,\ \ p(\mathbf{z}|\mathbf{h}_1)\,p(\mathbf{h}_2|\mathbf{h}_1)||_{\mathcal{T}}\right] \\
&= \mathcal{J}(\mathbf{z};\,\mathbf{h}_2\,|\,\mathbf{h}_1)
\end{aligned} \tag{A2}$$

Combining Equations (A1) and (A2) yields the inequality:

$$\mathcal{J}(\mathbf{z};\,(\mathbf{h}_1,\mathbf{h}_2)) \leq \mathcal{J}(\mathbf{z};\,\mathbf{h}_1) + \mathcal{J}(\mathbf{z};\,\mathbf{h}_2\,|\,\mathbf{h}_1) \tag{A3}$$

Next, we use Equation (A3) to prove the general statement for all $k \geq 1$. By writing:

$$\mathcal{J}(\mathbf{z};\,(\mathbf{h}_1,\ldots,\mathbf{h}_k)) \leq \mathcal{J}(\mathbf{z};\,\mathbf{h}_k\,|\,(\mathbf{h}_1,\ldots,\mathbf{h}_{k-1})) + \mathcal{J}(\mathbf{z};\,(\mathbf{h}_1,\ldots,\mathbf{h}_{k-1}))$$

Repeating the same inequality on the last term on the right-hand side yields the statement of the theorem.

Appendix C. Proof of Proposition 1

By the triangle inequality:

$$\begin{aligned}
\mathcal{J}(\mathbf{x};\,\mathbf{z}\,|\,\mathbf{y}) &= \mathbb{E}_{\mathbf{y}}||p(\mathbf{x}|\mathbf{y})\cdot p(\mathbf{z}|\mathbf{y})\,,\ \ p(\mathbf{x},\mathbf{z}|\mathbf{y})||_{\mathcal{T}} \\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}||p(\mathbf{z}|\mathbf{y})\,,\ \ p(\mathbf{z}|\mathbf{x},\mathbf{y})||_{\mathcal{T}} \\
&\leq \mathbb{E}_{\mathbf{x},\mathbf{y}}||p(\mathbf{z}|\mathbf{y})\,,\ \ p(\mathbf{z})||_{\mathcal{T}} + \mathbb{E}_{\mathbf{x},\mathbf{y}}||p(\mathbf{z})\,,\ \ p(\mathbf{z}|\mathbf{x},\mathbf{y})||_{\mathcal{T}} \\
&= \mathbb{E}_{\mathbf{y}}||p(\mathbf{z}|\mathbf{y})\,,\ \ p(\mathbf{z})||_{\mathcal{T}} + \mathbb{E}_{\mathbf{x},\mathbf{y}}||p(\mathbf{z})\,,\ \ p(\mathbf{z}|\mathbf{x},\mathbf{y})||_{\mathcal{T}} \\
&= \mathcal{J}(\mathbf{y};\,\mathbf{z}) + \mathcal{J}(\mathbf{z};\,(\mathbf{x},\mathbf{y}))
\end{aligned}$$

Therefore:

$$\mathcal{J}(\mathbf{z};\,(\mathbf{x},\mathbf{y})) \geq \mathcal{J}(\mathbf{x};\,\mathbf{z}\,|\,\mathbf{y}) - \mathcal{J}(\mathbf{y};\,\mathbf{z})$$

Combining this with the following chain rule of Theorem 2:

$$\mathcal{J}(\mathbf{z};\,(\mathbf{x},\mathbf{y})) \leq \mathcal{J}(\mathbf{x};\,\mathbf{z}\,|\,\mathbf{y}) + \mathcal{J}(\mathbf{y};\,\mathbf{z})$$

yields:

$$\Big|\mathcal{J}(\mathbf{z};\,(\mathbf{x},\mathbf{y})) - \mathcal{J}(\mathbf{x};\,\mathbf{z}\,|\,\mathbf{y})\Big| \leq \mathcal{J}(\mathbf{y};\,\mathbf{z})$$

Or equivalently:

$$\Big|\mathcal{J}(\mathbf{x};\,(\mathbf{y},\mathbf{z})) - \mathcal{J}(\mathbf{x};\,\mathbf{z}\,|\,\mathbf{y})\Big| \leq \mathcal{J}(\mathbf{x};\,\mathbf{y}) \tag{A4}$$

To prove the other inequality, we use Lemma 2. We have:

$$\mathcal{J}(\mathbf{x};\,\mathbf{y}) \leq \mathcal{J}(\mathbf{x};\,(\mathbf{y},\mathbf{z})) \leq \mathcal{J}(\mathbf{x};\,\mathbf{y}) + \mathcal{J}(\mathbf{x};\,\mathbf{z}\,|\,\mathbf{y}),$$

where the first inequality follows from Lemma 2 and the second inequality follows from the chain rule. Thus, we obtain the desired bound:

$$\Big|\mathcal{J}(\mathbf{x};\,(\mathbf{y},\mathbf{z})) - \mathcal{J}(\mathbf{x};\,\mathbf{y})\Big| \leq \mathcal{J}(\mathbf{x};\,\mathbf{z}\,|\,\mathbf{y}) \tag{A5}$$

Both Equations (A4) and (A5) imply that the chain rule is tight. More precisely, the inequality can be made arbitrarily close to an equality when one of the two terms in the upper bound is chosen to be arbitrarily close to zero.

Appendix D. Proof of Theorem 3

We will use the following fact:

Fact 1. *Let $f : \mathcal{X} \to [0,1]$ be a function with a bounded range in the interval $[0,1]$. Let $p_1(x)$ and $p_2(x)$ be two different probability measures defined on the same space $\mathcal{X}$. Then:*

$$\left|\mathbb{E}_{\boldsymbol{x}\sim p_1(x)} f(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_2(x)} f(\boldsymbol{x})\right| \leq ||p_1(x)\,,\;\; p_2(x)||_{\mathcal{T}}$$

First Setting: We first consider the following scenario. Suppose a learning algorithm $\mathcal{L}$ produces a hypothesis $\mathbf{h} \in \mathcal{H}$ from some marginal distribution $p(h)$ *independently* of the training sample $\mathbf{s}$. Afterwards, $\mathcal{L}$ produces a second hypothesis $\mathbf{k} \in \mathcal{K}$ according to $p(k\,|\,\mathbf{h},\mathbf{s})$. In other words, $\mathbf{k}$ depends on both $\mathbf{h}$ and $\mathbf{s}$ but the latter two random variables are independent of each other. Under this scenario, we have:

$$\mathcal{J}(\hat{\mathbf{z}};\,(\mathbf{h},\mathbf{k})) = \mathcal{J}(\hat{\mathbf{z}};\,\mathbf{k}\,|\,\mathbf{h}),$$

where the equality follows from the chain rule in Theorem 1, the statement of Proposition 1, and the fact that $\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{h}) = 0$.

The conditional variational information is written as:

$$\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{k}\,|\,\mathbf{h}) = \mathbb{E}_{\mathbf{h}}||p(\hat{\mathbf{z}})\cdot p(\mathbf{k}|\mathbf{h})\,,\;\; p(\hat{\mathbf{z}},\mathbf{k}|\mathbf{h})||_{\mathcal{T}},$$

where we used the fact that $p(\hat{\mathbf{z}}|\mathbf{h}) = p(\hat{\mathbf{z}})$. By marginalization:

$$p(\mathbf{k}|\mathbf{h}) = \mathbb{E}_{\hat{\mathbf{z}}'|\mathbf{h}}[p(\mathbf{k}|\hat{\mathbf{z}}',\mathbf{h})] = \mathbb{E}_{\hat{\mathbf{z}}'\sim p(z)}[p(\mathbf{k}|\hat{\mathbf{z}}',\mathbf{h})]$$

Similarly:

$$p(\hat{\mathbf{z}},\mathbf{k}|\mathbf{h}) = p(\hat{\mathbf{z}}|\mathbf{h})\cdot p(\mathbf{k}|\hat{\mathbf{z}},\mathbf{h}) = p(\hat{\mathbf{z}})\cdot p(\mathbf{k}|\hat{\mathbf{z}},\mathbf{h})$$

Therefore:

$$\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{k}\,|\,\mathbf{h}) = \mathbb{E}_{\mathbf{h}}\mathbb{E}_{\hat{\mathbf{z}}}||\mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k}|\hat{\mathbf{z}}',\mathbf{h})]\,,\;\; p(\mathbf{k}|\hat{\mathbf{z}},\mathbf{h})||_{\mathcal{T}}$$

Next, we note that since $\mathbf{h}$ is independent of the sample $\mathbf{s}$, the variational information between $\hat{\mathbf{z}} \sim \mathbf{s}$ and $\mathbf{k} \in \mathcal{K}$ can be bounded using Theorem 6. This follows because $\mathbf{h}$ is selected independently of the sample $\mathbf{s}$, and, hence, the i.i.d. property of the observations $\mathbf{z}_i$ continues to hold. Therefore, we obtain:

$$\mathbb{E}_{\mathbf{h}}\mathbb{E}_{\hat{\mathbf{z}}}||\mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k}|\hat{\mathbf{z}}',\mathbf{h})]\,,\;\; p(\mathbf{k}|\hat{\mathbf{z}},\mathbf{h})||_{\mathcal{T}} \leq \sqrt{\frac{\log|\mathcal{K}|}{2m}} \tag{A6}$$

Because $p(\mathbf{k}|\hat{\mathbf{z}},\mathbf{h})$ is arbitrary in our derivation, the above bound holds for any distribution of observations $p(z)$, any distribution $p(h)$, and any family of conditional distributions $p(k|\hat{\mathbf{z}},\mathbf{h})$.

Original Setting: Next, we return to the original setting where both $\mathbf{h} \in \mathcal{H}$ and $\mathbf{k} \in \mathcal{K}$ are chosen *according* to the training sample $\mathbf{s}$. We have:

$$\begin{aligned}
\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{k} \,|\, \mathbf{h}) &= \mathbb{E}_{\mathbf{h}} || p(\hat{\mathbf{z}}|\mathbf{h}) \cdot p(\mathbf{k}|\mathbf{h}) \,,\; p(\hat{\mathbf{z}}, \mathbf{k}|\mathbf{h}) ||_{\mathcal{T}} \\
&= \mathbb{E}_{\mathbf{h},\hat{\mathbf{z}}} || p(\mathbf{k}|\mathbf{h}) \,,\; p(\mathbf{k}|\hat{\mathbf{z}}, \mathbf{h}) ||_{\mathcal{T}} \\
&= \mathbb{E}_{\mathbf{h},\hat{\mathbf{z}}} || \mathbb{E}_{\hat{\mathbf{z}}'|\mathbf{h}}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] \,,\; p(\mathbf{k}|\hat{\mathbf{z}}, \mathbf{h}) ||_{\mathcal{T}} \\
&\leq \mathbb{E}_{\mathbf{h},\hat{\mathbf{z}}} || \mathbb{E}_{\hat{\mathbf{z}}'|\mathbf{h}}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] \,,\; \mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] ||_{\mathcal{T}} + \mathbb{E}_{\mathbf{h},\hat{\mathbf{z}}} || \mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] \,,\; p(\mathbf{k}|\hat{\mathbf{z}}, \mathbf{h}) ||_{\mathcal{T}}
\end{aligned} \tag{A7}$$

In the last line, we used the triangle inequality.

Next, we would like to bound the first term. Using the fact that the total variation distance is related to the ℓ_1 distance by $||p \,,\; q||_{\mathcal{T}} = \frac{1}{2}||p - q||_1$, we have:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{h},\hat{\mathbf{z}}} || \mathbb{E}_{\hat{\mathbf{z}}'|\mathbf{h}}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] \,,\; \mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] ||_{\mathcal{T}} \\
&= \mathbb{E}_{\mathbf{h}} \, || \mathbb{E}_{\hat{\mathbf{z}}'|\mathbf{h}}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] \,,\; \mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] ||_{\mathcal{T}} \\
&= \frac{1}{2} \, \mathbb{E}_{\mathbf{h}} \sum_{k \in \mathcal{K}} \left| \mathbb{E}_{\hat{\mathbf{z}}'|\mathbf{h}}[p(\mathbf{k} = k|\hat{\mathbf{z}}', \mathbf{h})] - \mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k} = k|\hat{\mathbf{z}}', \mathbf{h})] \right| \\
&\leq \frac{1}{2} \sum_{\mathbf{k} \in \mathcal{K}} \mathbb{E}_{\mathbf{h}} || p(\hat{\mathbf{z}}'|\mathbf{h}) \,,\; p(\hat{\mathbf{z}}') ||_{\mathcal{T}} \\
&= \frac{1}{2} \sum_{\mathbf{k} \in \mathcal{K}} \mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) = \frac{|\mathcal{K}|}{2} \, \mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h})
\end{aligned} \tag{A8}$$

Here, the inequality follows from Fact 1.

Next, we bound the second term in Equation (A7). Using Fact 1 and our earlier result in Equation (A6):

$$\begin{aligned}
&\mathbb{E}_{\mathbf{h},\hat{\mathbf{z}}} || \mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] \,,\; p(\mathbf{k}|\hat{\mathbf{z}}, \mathbf{h}) ||_{\mathcal{T}} \\
&\leq \mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) + \mathbb{E}_{\mathbf{h}} \mathbb{E}_{\hat{\mathbf{z}}} || \mathbb{E}_{\hat{\mathbf{z}}'}[p(\mathbf{k}|\hat{\mathbf{z}}', \mathbf{h})] \,,\; p(\mathbf{k}|\hat{\mathbf{z}}, \mathbf{h}) ||_{\mathcal{T}} \\
&\leq \mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) + \sqrt{\frac{\log |\mathcal{K}|}{2m}}
\end{aligned} \tag{A9}$$

Combining all results in Equations (A7)–(A9):

$$\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{k} \,|\, \hat{\mathbf{z}}) \leq \left[1 + \frac{|\mathcal{K}|}{2} \right] \mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) + \sqrt{\frac{\log |\mathcal{K}|}{2m}} \tag{A10}$$

This along with the chain rule imply the statement of the theorem.

Appendix E. Proof of Proposition 3

Let $I(\mathbf{x};\mathbf{y})$ denote the mutual information between $\mathbf{x}$ and $\mathbf{y}$ and let $H(\mathbf{x})$ denote the Shannon entropy of the random variable $\mathbf{x}$ measured in nats (i.e., using natural logarithms). As before, we write $\mathbf{s} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)$. We have:

$$\begin{aligned} I(\mathbf{s};(\mathbf{h},\mathbf{k})) &= H(\mathbf{s}) - H(\mathbf{s} \mid \mathbf{h},\mathbf{k}) \\ &= \sum_{i=1}^{m} H(\mathbf{z}_i) - \sum_{i=1}^{m} H(\mathbf{z}_i|\mathbf{h},\mathbf{k},\mathbf{z}_1,\ldots,\mathbf{z}_{i-1}) \\ &\geq \sum_{i=1}^{m} H(\mathbf{z}_i) - H(\mathbf{z}_i|\mathbf{h},\mathbf{k}) = mI(\hat{\mathbf{z}};\mathbf{h},\mathbf{k}) \end{aligned}$$

The second line is the chain rule for entropy and the third lines follows from the fact that conditioning reduces entropy. We obtain:

$$I(\hat{\mathbf{z}};\mathbf{h},\mathbf{k}) \leq \frac{I(\mathbf{s};(\mathbf{h},\mathbf{k}))}{m}$$

By Pinsker's inequality:

$$\mathcal{J}(\hat{\mathbf{z}};\,(\mathbf{h},\mathbf{k})) \leq \sqrt{\frac{I(\hat{\mathbf{z}};(\mathbf{h},\mathbf{k}))}{2}} \leq \sqrt{\frac{I(\mathbf{s};(\mathbf{h},\mathbf{k}))}{2m}}$$

Using the chain rule for mutual information:

$$\begin{aligned} \mathcal{J}(\hat{\mathbf{z}};\,(\mathbf{h},\mathbf{k})) &\leq \sqrt{\frac{I(\mathbf{s};(\mathbf{h},\mathbf{k}))}{2m}} = \sqrt{\frac{I(\mathbf{s};\mathbf{h}) + I(\mathbf{s};\mathbf{k}|\mathbf{h})}{2m}} \\ &\leq \sqrt{\frac{I(\mathbf{s};\mathbf{h}) + H(\mathbf{k})}{2m}} \leq \sqrt{\frac{I(\mathbf{s};\mathbf{h}) + \log|\mathbf{k}|}{2m}} \end{aligned}$$

The desired bound follows by applying the same proof technique of Theorem 4 on the last uniform generalization bound, and using the fact that $\log 3 < 2$.

Appendix F. Proof of Proposition 4

Before we prove the statement of the theorem, we begin with the following lemma:

Lemma A1. *Let the observation space $\mathcal{Z}$ be the interval $[0,1]$, where $p(z)$ is continuous in $[0,1]$. Let $\boldsymbol{h} \subseteq \boldsymbol{s} : |\boldsymbol{h}| = k$ be a set of k examples picked at random without replacement from the training sample $\boldsymbol{s}$. Then $\mathcal{J}(\hat{\mathbf{z}};\,\boldsymbol{h}) = \frac{k}{m}$.*

Proof. First, we note that $p(\hat{\mathbf{z}}|\mathbf{h})$ is a *mixture* of two distributions: one that is uniform in $\mathbf{h}$ with probability k/m, and the original distribution $p(z)$ with probability $1 - k/m$. By Jensen's inequality, we have $\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{h}) \leq k/m$. Second, let the parametric loss be $l(z;\mathbf{h}) = \mathbb{I}\{z \in \mathbf{h}\}$. Then, $|R_{gen}(\mathcal{L})| = \frac{k}{m}$. By Theorem 2, we have $\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{h}) \geq |R_{gen}(\mathcal{L})| = k/m$. Both bounds imply the statement of the lemma. □

Now, we prove Proposition 4. Consider the setting where $\mathcal{Z} = [0,\,1]$ and suppose that the observations $\mathbf{z} \in \mathcal{Z}$ have a continuous marginal distribution. Because t is a rational number, let the sample size m be chosen such that $k = t\,m$ is an integer.

Let $\mathbf{s} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ be the training set, and let the hypothesis $\mathbf{h}$ be given by $\mathbf{h} = \{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$ with some probability $\delta > 0$ and $\mathbf{h} = \{\}$ otherwise. Here, the k instances $\mathbf{z}_i \in \mathbf{h}$ are picked uniformly at random

without replacement from the sample $\mathbf{s}$. To determine the variational information between $\hat{\mathbf{z}}$ and $\mathbf{h}$, we consider the two cases:

1. If $\mathbf{h} \neq \{\}$, then $||p(\hat{\mathbf{z}})\,,\ \ p(\hat{\mathbf{z}}|\mathbf{h})||_{\mathcal{T}} = t$ as proved in Lemma 1. This happens with probability δ by design.
2. Otherwise, $p(\hat{\mathbf{z}}|\mathbf{h}) = p(\hat{\mathbf{z}})$. Thus: $||p(\hat{\mathbf{z}})\,,\ \ p(\hat{\mathbf{z}}|\mathbf{h})||_{\mathcal{T}} = 0$.

So, by combining the two cases above, we deduce that:

$$\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) = \mathbb{E}_{\mathbf{h}}\, ||p(\hat{\mathbf{z}})\,,\ \ p(\hat{\mathbf{z}} \mid \mathbf{h})||_{\mathcal{T}} = t\ \delta.$$

Therefore, $\mathcal{L}$ generalizes uniformly with the rate $t\delta$. Next, let the parametric loss be given by $l(z\,;\ \mathbf{h}) = \mathbb{I}\{z \in \mathbf{h}\}$. With this loss:

$$p\Big\{\big|R_{\mathbf{s}}(\mathbf{h}) - R(\mathbf{h})\big| = t\Big\} = \delta = \frac{\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h})}{t},$$

which is the statement of the proposition.

Appendix G. Proof of Theorem 5

Because $\mathcal{Z}$ is countable, we will assume without loss of generality that $\mathcal{Z} = \{1, 2, 3, \ldots, \ldots\}$, and we will write $p_z = p(\hat{\mathbf{z}} = z)$ to denote the marginal distribution of observations. Since all lazy learners are equivalent, we will look into the lazy learner whose hypothesis $\mathbf{h}$ is equal to the training sample $\mathbf{s}$ itself up to a permutation. Let m_z denote the number of times $z \in \mathcal{Z}$ was observed in the training sample. Note that $p(\hat{\mathbf{z}} = z|\mathbf{h}) = p_{\mathbf{s}}(z)$, and so $\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) = \mathbb{E}_{\mathbf{s}}\, ||p(z),\ p_{\mathbf{s}}(z)||_{\mathcal{T}}$.

We have:

$$p(\mathbf{h}) = p(\mathbf{s}) = \binom{m}{m_1,\, m_2,\, \ldots}\, p_1^{m_1}\, p_2^{m_2} \cdots$$

Using the relation $||p\,,\ \ q||_{\mathcal{T}} = \frac{1}{2}||p - q||_1$ for any two probability distributions p and q, we obtain:

$$\mathbb{E}_{\mathbf{h}}\, ||p(\hat{\mathbf{z}}) - p(\hat{\mathbf{z}}|\mathbf{h})||_1 = \sum_{k \geq 1\,:\, m_1 + m_2 + .. = m} \binom{m}{m_1, m_2, \ldots} \times p_1^{m_1} p_2^{m_2} \cdots \left|\frac{m_k}{m} - p_k\right|$$

For the inner summation, we write:

$$\begin{aligned}
&\sum_{m_1 + m_2 + \ldots = m} \binom{m}{m_1,\, m_2,\, \ldots}\, p_1^{m_1}\, p_2^{m_2} \cdots \left|\frac{m_k}{m} - p_k\right| \\
&= \sum_{s=0}^{m} \binom{m}{s}\, p_k^s \left|\frac{m_k}{m} - p_k\right| \sum_{m_1 + \ldots + m_{k-1} + m_{k+1} + \ldots = m - s} \binom{m-s}{m_1, \ldots, m_{k-1},\, m_{k+1}, \ldots} \times p_1^{m_1} \cdots p_{k-1}^{m_{k-1}}\, p_{k+1}^{m_{k+1}} \cdots
\end{aligned}$$

Using the multinomial series, we simplify the right-hand side into:

$$\sum_{s=0}^{m} \binom{m}{s}\, p_k^s\, (1 - p_k)^{m-s} \left|\frac{s}{m} - p_k\right|$$

Now, we use *De Moivre's formula* for the mean deviation of the binomial random variable (see the proof of Example 1). This gives us:

$$\begin{aligned}
&\sum_{m_1+m_2+\ldots=m} \binom{m}{m_1,\, m_2,\, \ldots} p_1^{m_1}\, p_2^{m_2} \cdots \left|\frac{s}{m} - p_k\right| \\
&= \sum_{s=0}^{m} \binom{m}{s} p_k^s\, (1-p_k)^{m-s} \left|\frac{s}{m} - p_k\right| \\
&= \frac{2}{m}\, (1-p_k)^{(1-p_k)m} p_k^{1+mp_k} \frac{m!}{(p_k m)!\, ((1-p_k)m-1)!}
\end{aligned}$$

Using *Stirling's approximation* to the factorial [17], we obtain the simple asymptotic expression:

$$\sum_{m_1+m_2+\ldots=m} \binom{m}{m_1,\, m_2,\, \ldots} p_1^{m_1}\, p_2^{m_2} \cdots \left|\frac{m_k}{m} - p_k\right| \;\sim\; \sqrt{\frac{2p_k(1-p_k)}{\pi m}}$$

Plugging this into the earlier expression for $\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h})$ yields:

$$\begin{aligned}
\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) &\sim \frac{1}{2} \sum_{k=1,2,3,\ldots} \sqrt{\frac{2p_k(1-p_k)}{\pi m}} \\
&= \sqrt{\frac{\mathbf{Ess}\left[\mathcal{Z};\, p(z)\right] - 1}{2\pi m}}
\end{aligned}$$

Due to the tightness of the Stirling approximation, the asymptotic expression for the variational information is tight. Because $\mathcal{J}(\hat{\mathbf{z}};\, \mathbf{h}) = \mathbb{E}_{\mathbf{s}}\, ||p(z),\, p_{\mathbf{s}}(z)||_{\mathcal{T}}$, we deduce that:

$$\mathbb{E}_{\mathbf{s}}\, ||p(z),\, p_{\mathbf{s}}(z)||_{\mathcal{T}} \sim \sqrt{\frac{\mathbf{Ess}\left[\mathcal{Z};\, p(z)\right] - 1}{2\,\pi\, m}},$$

which provides the asymptotic rate of convergence of an empirical probability mass function to the true distribution.

Appendix H. Proof of Proposition 5

First, we note that for any two adjacent samples $\mathbf{s}$ and $\mathbf{s}'$ and any $\mathcal{O} \subseteq \mathcal{H}$, we have in the differential privacy setting:

$$p(\mathbf{h} \in \mathcal{O}|\mathbf{s}) - p(\mathbf{h} \in \mathcal{O}|\mathbf{s}') \leq (e^{\epsilon} - 1)\; p(\mathbf{h} \in \mathcal{O}|\mathbf{s}') + \delta$$

Similarly, we have:

$$\begin{aligned}
p(\mathbf{h} \in \mathcal{O}|\mathbf{s}) - p(\mathbf{h} \in \mathcal{O}|\mathbf{s}') &\geq (e^{-\epsilon} - 1)\; p(\mathbf{h} \in \mathcal{O}|\mathbf{s}') - e^{-\epsilon}\delta \\
&= -\Big[(1 - e^{-\epsilon})\; p(\mathbf{h} \in \mathcal{O}|\mathbf{s}') + e^{-\epsilon}\delta\Big] \\
&\geq -e^{\epsilon}\; \Big[(1 - e^{-\epsilon})\; p(\mathbf{h} \in \mathcal{O}|\mathbf{s}') + e^{-\epsilon}\delta\Big] \\
&= -\Big[(e^{\epsilon} - 1) p(\mathbf{h} \in \mathcal{O}|\mathbf{s}') + \delta\Big]
\end{aligned}$$

Both results imply that:

$$\begin{aligned}\left|p(\mathbf{h}\in\mathcal{O}|\mathbf{s})-p(\mathbf{h}\in\mathcal{O}|\mathbf{s}')\right| &\leq (e^{\epsilon}-1)p(\mathbf{h}\in\mathcal{O}|\mathbf{s}')+\delta\\ &\leq e^{\epsilon}-1+\delta\end{aligned} \tag{A11}$$

We write:

$$\begin{aligned}\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{h}) &= \mathbb{E}_{\hat{\mathbf{z}}}\;||p(\mathbf{h}|\hat{\mathbf{z}})\,,\;\;p(\mathbf{h})||_{\mathcal{T}}\\ &= \frac{1}{2}\mathbb{E}_{\hat{\mathbf{z}}}\big|\big|\mathbb{E}_{\hat{\mathbf{z}}'}\left[p(\mathbf{h}|\hat{\mathbf{z}})-p(\mathbf{h}|\hat{\mathbf{z}}')\right]\big|\big|_1\\ &\leq \frac{1}{2}\mathbb{E}_{\hat{\mathbf{z}},\hat{\mathbf{z}}'}\big|\big|p(\mathbf{h}|\hat{\mathbf{z}})-p(\mathbf{h}|\hat{\mathbf{z}}')\big|\big|_1\end{aligned}$$

The last inequality follows by convexity. Next, let $\mathbf{s}_{m-1}$ be a sample that contains $m-1$ observations drawing i.i.d. from $p(z)$. Then:

$$\begin{aligned}\mathcal{J}(\hat{\mathbf{z}};\,\mathbf{h}) &\leq \frac{1}{2}\mathbb{E}_{\hat{\mathbf{z}},\hat{\mathbf{z}}'}\big|\big|\mathbb{E}_{\mathbf{s}_{m-1}}\left[p(\mathbf{h}|\hat{\mathbf{z}},\mathbf{s}_{m-1})-p(\mathbf{h}|\hat{\mathbf{z}}',\mathbf{s}_{m-1})\right]\big|\big|_1\\ &\leq \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{s}'}\big|\big|p(\mathbf{h}|\mathbf{s})-p(\mathbf{h}|\mathbf{s}')\big|\big|_1,\end{aligned}$$

where $\mathbf{s},\mathbf{s}'$ are two adjacent samples. Finally, we use Equation (A11) to arrive at the statement of the proposition.

Appendix I. Proof of Theorem 8

The proof is similar to the classical VC argument. Given a fixed hypothesis space $\mathcal{H}$, a fixed domain $\mathcal{Z}$, and a 0–1 loss function $l:\mathcal{H}\times\mathcal{Z}\to\{0,1\}$, let $\mathbf{s}=\{\mathbf{z}_1,\ldots,\mathbf{z}_m\}$ be a training sample that comprises of m i.i.d. observations. Define the *restriction* of $\mathcal{H}$ to $\mathbf{s}$ by:

$$\mathcal{F}_{\mathbf{s}}=\Big\{l(\mathbf{z}_1,\,h),\ldots,l(\mathbf{z}_m,\,h):\;\;h\in\mathcal{H}\Big\}$$

In other words, $\mathcal{F}_{\mathbf{s}}$ is the set of all possible realizations of the 0–1 loss for the elements in $\mathbf{s}$ by hypotheses in $\mathcal{H}$. We can introduce an *equivalence relation* between the elements of $\mathcal{H}$ w.r.t. the sample $\mathbf{s}$. Specifically, we say that for $h',h''\in\mathcal{H}$, we have $h'\equiv_{\mathbf{s}}h''$ if and only if:

$$\big(l(\mathbf{z}_1,h'),\ldots,l(\mathbf{z}_m,h')\big)=\big(l(\mathbf{z}_1,h''),\ldots,l(\mathbf{z}_m,h'')\big)$$

It is trivial to see that this defines an equivalence relation; i.e., it is reflexive, symmetric, and transitive. Let the set of equivalence classes w.r.t. $\mathbf{s}$ be denoted $\mathcal{H}_{\mathbf{s}}$. Note that we have a one-to-one correspondence between the members of $\mathcal{F}_{\mathbf{s}}$ and the members of $\mathcal{H}_{\mathbf{s}}$. Moreover, $\mathcal{H}_{\mathbf{s}}$ is a *partitioning* of $\mathcal{H}$.

We use the standard twin-sample trick where we have $\mathbf{s}_2=\mathbf{s}\cup\mathbf{s}'\in\mathcal{Z}^{2m}$ and $\mathcal{L}$ learns based on $\mathbf{s}$ only. For any fixed $h\in\mathcal{H}$, let $f:\mathcal{H}\times\mathcal{Z}\to[0,1]$ be an arbitrary loss function, which can be different from the loss l that is optimized during the training. A Hoeffding bound for sampling without replacement [41] states that:

$$p\Big\{\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}[f(\mathbf{z},h)]-\mathbb{E}_{\mathbf{z}\sim\mathbf{s}_2}[f(\mathbf{z},h)]\Big|\;\geq\epsilon\Big\}\leq 2\exp\{-2\epsilon^2 m\} \tag{A12}$$

Hence:

$$\begin{aligned}
&p\Big\{\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}[f(\mathbf{z},h)]-\mathbb{E}_{\mathbf{z}\sim\mathbf{s}'}[f(\mathbf{z},h)]\Big| \ \geq \epsilon\Big\}\\
&\quad\leq p\Big\{\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}[f(\mathbf{z},h)]-\mathbb{E}_{\mathbf{z}\sim\mathbf{s}_2}[f(\mathbf{z},h)]\Big| \ \geq \frac{\epsilon}{2}\Big\}+p\Big\{\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}'}[f(\mathbf{z},h)]-\mathbb{E}_{\mathbf{z}\sim\mathbf{s}_2}[f(\mathbf{z},h)]\Big| \ \geq \frac{\epsilon}{2}\Big\}\\
&\quad\leq 4\exp\{-(1/2)\epsilon^2 m\}
\end{aligned}$$

This happens for a hypothesis $h\in\mathcal{H}$ that is fixed independently of the random split of $\mathbf{s}_2$ into training and ghost samples. When h is selected according to the random split of $\mathbf{s}_2$, then we need to employ the union bound.

For any subset $H\subseteq\mathcal{H}$, let $\min(H)$ be the least element in H according to $\preceq$. Let $\mathcal{H}_{\mathbf{s}}$ be as defined previously and write $H_{min}(\mathbf{s})=\{\min(H_k):H_k\in\mathcal{H}_{\mathbf{s}}\}$. Then, it is easy to observe that the ERM learning rule of Theorem 2 must select one of the hypotheses in $H_{min}(\mathbf{s}_2)$ regardless of the split $\mathbf{s}_2=\mathbf{s}\cup\mathbf{s}'$. This holds because $\mathcal{H}_{\mathbf{s}_2}$ is a *coarser* partitioning of $\mathcal{H}$ than $\mathcal{H}_{\mathbf{s}}$. In other words, every member of $\mathcal{H}_{\mathbf{s}}$ is a union of some finite number of members of $\mathcal{H}_{\mathbf{s}_2}$. By the well-ordering property, the " least" element among the empirical risk minimizers must be in $H_{min}(\mathbf{s}_2)$.

Hence, there is, at most, $\tau_{\mathcal{H}}(2m)$ possible hypotheses given $\mathbf{s}_2$, where $\tau_{\mathcal{H}}(m)$ is the growth function (sometimes referred to as the shattering coefficient), and those hypotheses can be fixed independently of the random splitting of $\mathbf{s}_2$ into a training sample $\mathbf{s}$ and a ghost sample $\mathbf{s}'$.

Consequently, we have by the union bound:

$$\begin{aligned}
&p\Big\{\sup_{h\in H_{min}(\mathbf{s}\cup\mathbf{s}')}\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}[f(\mathbf{z},h)]-\mathbb{E}_{\mathbf{z}\sim\mathbf{s}'}[f(\mathbf{z},h)]\Big| \ \geq \epsilon\Big\}\\
&\quad\leq 4\tau_{\mathcal{H}}(2m)\exp\{-\frac{\epsilon^2 m}{2}\}\leq 4\Big(\frac{2em}{d}\Big)^d\exp\{-\frac{\epsilon^2 m}{2}\},
\end{aligned}$$

where d is the VC dimension of $\mathcal{H}$. Finally, to bound the generalization risk in expectation, we use Lemma A.4 in [13], which implies that if $m\geq d$:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s},\mathbf{s}'}\Big[\sup_{h\in H_{min}(\mathbf{s}\cup\mathbf{s}')}\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}[f(\mathbf{z},h)]-\mathbb{E}_{\mathbf{z}\sim\mathbf{s}'}[f(\mathbf{z},h)]\Big|\Big]\\
&\leq\sqrt{\frac{2}{m}}\Big(2+\sqrt{\log 2+d\log\frac{2em}{d}}\Big)\\
&\leq\sqrt{\frac{2}{m}}\Big(2+\sqrt{1+d\log\frac{2em}{d}}\Big)\leq\frac{3+\sqrt{1+d\log\frac{2em}{d}}}{\sqrt{m}}
\end{aligned}$$

Writing $\hat{\mathbf{h}}$ for the *least* empirical risk minimizer w.r.t. the training sample $\mathbf{s}$:

$$\begin{aligned}
R_{gen}(\mathcal{L}) &= \mathbb{E}_{\mathbf{s}}\Big[\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}\left[f(\mathbf{z},\hat{\mathbf{h}})\right] - \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[f(\mathbf{z},\hat{\mathbf{h}})\right]\Big] \\
&\leq \mathbb{E}_{\mathbf{s}}\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}\left[f(\mathbf{z},\hat{\mathbf{h}})\right] - \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[f(\mathbf{z},\hat{\mathbf{h}})\right]\Big| \\
&= \mathbb{E}_{\mathbf{s}}\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}\left[f(\mathbf{z},\hat{\mathbf{h}})\right] - \mathbb{E}_{\mathbf{s}'}\mathbb{E}_{\mathbf{z}\sim\mathbf{s}'}\left[f(\mathbf{z},\hat{\mathbf{h}})\right]\Big| \\
&\leq \mathbb{E}_{\mathbf{s},\mathbf{s}'}\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}\left[f(\mathbf{z},\hat{\mathbf{h}})\right] - \mathbb{E}_{\mathbf{z}\sim\mathbf{s}'}\left[f(\mathbf{z},\hat{\mathbf{h}})\right]\Big| \\
&\leq \mathbb{E}_{\mathbf{s},\mathbf{s}'}\sup_{h\in H_{min}(\mathbf{s}\cup\mathbf{s}')}\Big|\mathbb{E}_{\mathbf{z}\sim\mathbf{s}}\left[f(\mathbf{z},h)\right] - \mathbb{E}_{\mathbf{z}\sim\mathbf{s}'}\left[f(\mathbf{z},h)\right]\Big| \\
&\leq \frac{3+\sqrt{1+d\log\frac{2em}{d}}}{\sqrt{m}}
\end{aligned}$$

Because this bound in expectation holds for any single loss $f : H \times \mathcal{Z} \to [0,1]$, it holds for the following loss function:

$$l^{\star}(z,h) = \mathbb{I}\big\{p(z\in\mathbf{s}|h) > p(z\in\mathbf{s})\big\},$$

which is a deterministic 0–1 loss function of h that assigns to $z \in \mathcal{Z}$ the value 1 if and only if our knowledge of h increases the probability that z belongs to the training sample. However, the generalization risk in expectation for the loss $l^{\star}$ is equal to the variational information $\mathcal{J}(\hat{\mathbf{h}};\hat{\mathbf{z}})$ as shown in the proof of Theorem 2. Hence, we have the bound stated in the theorem:

$$\mathcal{J}(\hat{\mathbf{h}};\hat{\mathbf{z}}) \leq \frac{3+\sqrt{1+d\log\frac{2em}{d}}}{\sqrt{m}},$$

Because this is a distribution-free bound, we have:

$$C(\mathcal{L}) \leq \frac{3+\sqrt{1+d\log\frac{2em}{d}}}{\sqrt{m}}$$

Appendix J. Proof of Theorem 9

Let $\mathcal{X}^{\star} = \{x_1,\ldots,x_d\}$ be a set of d points in $\mathcal{X}$ that are shattered by hypotheses in $\mathcal{H}$. By definition, this implies that for any possible 0–1 labeling $I \in \{0,1\}^d$, there exists a hypothesis $h_I \in \mathcal{H}$ such that $(h_I(x_1),\ldots,h_I(x_d)) = I$.

Given an ERM learning rule $\mathcal{L}$ whose hypothesis is denoted $\hat{\mathbf{h}}_{\mathbf{s}}$, let $p(x)$ be the uniform distribution of instances over $\mathcal{X}^{\star}$ and define:

$$y(\mathbf{x}) = \arg\min_{\tilde{y}\in\{+1,-1\}} p_{\mathbf{s}}\Big\{\hat{\mathbf{h}}_{\mathbf{s}}(\mathbf{x}) = \tilde{y} \mid \mathbf{x}\notin\mathbf{s}\Big\}$$

In other words, $y(\mathbf{x})$ is the least probable class that is assigned by $\mathcal{L}$ to the instance $\mathbf{x}$ when $\mathbf{x}$ is unseen in the training sample. Let $p(z)$ with $\mathbf{z} = (\mathbf{x},\mathbf{y})$ denote the uniform distribution of instances over $\mathcal{X}^{\star}$ with $\mathbf{y}$ given by the labeling rule above.

By drawing a training sample $\mathbf{s} \in \mathcal{Z}^m$ of m i.i.d. observations from $p(z)$, our first task is to bound the expected number of *distinct* values in $\mathcal{X}^{\star}$ that are not observed in the training sample. Let:

$$E_i = \mathbb{I}\{x_i \notin \mathbf{s}\}$$

Then, the expected number of *distinct* values in $\mathcal{X}^\star$ that are not observed in the training sample $\mathbf{s}$ is:

$$\sum_{i=1}^{d} \mathbb{E}[E_i] = \sum_{i=1}^{d} \left(1 - \frac{1}{d}\right)^m = d\left(1 - \frac{1}{d}\right)^m$$

Here, we used the linearity of expectation, which holds even when the random variables are not independent. This shows that the expected *fraction* of instances in $\mathcal{X}^\star$ that are not seen in the sample $\mathbf{s}$ is $(1 - \frac{1}{d})^m$.

Next, given an ERM learning rule that outputs an empirical risk minimizer, the training error of this learning algorithm is zero because $\mathcal{X}^\star$ is shattered by $\mathcal{H}$. However, for any learning rule $\mathcal{L}$, the expected error rate on the unseen examples is, at least, $1/2$ by construction. Therefore, there exists a distribution $p(z)$ in which the generalization risk is, at least, $(1/2)(1 - 1/d)^m$.

By Theorem 2, the learning capacity is an upper bound on the maximum generalization risk across all distributions of observations and all parametric loss functions. Consequently:

$$C(\mathcal{L}) \geq \frac{1}{2}\left(1 - \frac{1}{d}\right)^m,$$

which is the statement of the theorem.

References

1. Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; Sridharan, K. Stochastic Convex Optimization. In Proceedings of the Annual Conference on Learning Theory, Montreal, QC, Canada, 18–21 June 2009.
2. Bartlett, P.L.; Jordan, M.I.; McAuliffe, J.D. Convexity, classification, and risk bounds. *J. Am. Stat. Assoc.* **2006**, *101*, 138–156. [CrossRef]
3. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef] [PubMed]
4. Blumer, A.; Ehrenfeucht, A.; Haussler, D.; Warmuth, M.K. Learnability and the Vapnik-Chervonenkis dimension. *JACM* **1989**, *36*, 929–965. [CrossRef]
5. McAllester, D. PAC-Bayesian stochastic model selection. *Mach. Learn.* **2003**, *51*, 5–21. [CrossRef]
6. Bousquet, O.; Elisseeff, A. Stability and generalization. *JMLR* **2002**, *2*, 499–526.
7. Bartlett, P.L.; Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR* **2002**, *3*, 463–482.
8. Kutin, S.; Niyogi, P. Almost-everywhere algorithmic stability and generalization error. In Proceedings of the Eighteenth conference on Uncertainty in Artificial Intelligence (UAI), Edmonton, AB, Canada, 1–4 August 2002.
9. Poggio, T.; Rifkin, R.; Mukherjee, S.; Niyogi, P. General conditions for predictivity in learning theory. *Nature* **2004**, *428*, 419–422. [CrossRef]
10. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley & Sons: New York, NY, USA, 1991.
11. Hardt, M.; Recht, B.; Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv* **2015**, arXiv:1509.01240.
12. Dwork, C.; Feldman, V.; Hardt, M.; Pitassi, T.; Reingold, O.; Roth, A. Preserving Statistical Validity in Adaptive Data Analysis. In Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC), Portland, OR, USA, 14–17 June 2015; pp. 117–126.
13. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: New York, NY, USA, 2014.

14. Raginsky, M.; Rakhlin, A.; Tsao, M.; Wu, Y.; Xu, A. Information-theoretic analysis of stability and bias of learning algorithms. In Proceedings of the 2016 IEEE Information Theory Workshop (ITW), Cambridge, UK, 11–14 September 2016; pp. 26–30.
15. Janson, S. Probability asymptotics: Notes on notation. *arXiv* **2011**, arXiv:1108.3924.
16. Tao, T. *Topics in Random Matrix Theory*; American Mathematical Society: Providence, RI, USA, 2012.
17. Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; Sridharan, K. Learnability, stability and uniform convergence. *JMLR* **2010**, *11*, 2635–2670.
18. Talagrand, M. Majorizing measures: The generic chaining. *Ann. Probab.* **1996**, *24*, 1049–1103. [CrossRef]
19. Audibert, J.Y.; Bousquet, O. Combining PAC-Bayesian and generic chaining bounds. *JMLR* **2007**, *8*, 863–889.
20. Xu, H.; Mannor, S. Robustness and generalization. *Mach. Learn.* **2012**, *86*, 391–423. [CrossRef]
21. Csiszár, I. A Class of Measures of Informativity of Observation Channels. *Period. Math. Hung.* **1972**, *2*, 191–213. [CrossRef]
22. Csiszár, I. Axiomatic Characterizations of Information Measures. *Entropy* **2008**, *10*, 261–273. [CrossRef]
23. Russo, D.; Zou, J. Controlling Bias in Adaptive Data Analysis Using Information Theory. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), Cadiz, Spain, 9–11 May 2016.
24. Bassily, R.; Moran, S.; Nachum, I.; Shafer, J.; Yehudayoff, A. Learners that Use Little Information. *PMLR* **2018**, *83*, 25–55.
25. Elkan, C. The foundations of cost-sensitive learning. In Proceedings of the IJCAI, Seattle, WA, USA, 4–10 August 2011.
26. Kull, M.; Flach, P. Novel decompositions of proper scoring rules for classification: score adjustment as precursor to calibration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Cham, Switzerland, 2015; pp. 68–85.
27. Robbins, H. A remark on Stirling's formula. *Am. Math. Mon.* **1955**, *62*, 26–29. [CrossRef]
28. Cortes, C.; Vapnik, V. Support vector machine. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
29. Wang, J.; Chen, Q.; Chen, Y. RBF kernel based support vector machine with universal approximation and its application. *ISNN* **2004**, *3173*, 512–517.
30. Downs, T.; Gates, K.E.; Masters, A. Exact simplification of support vector solutions. *JMLR* **2002**, *2*, 293–297.
31. Stigler, S.M. *The History of Statistics: The Measurement of Uncertainty before 1900*; Harvard University Press: Cambridge, MA, USA, 1986.
32. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Third Theory of Cryptography Conference (TCC 2006), New York, NY, USA, 4–7 March 2006; pp. 265–284.
33. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Theor. Comput. Sci.* **2013**, *9*, 211–407.
34. Koren, T.; Levy, K. Fast rates for exp-concave empirical risk minimization. In Proceedings of the NIPS 2015, Montreal, QC, Canada, 7–12 December, 2015; pp. 1477–1485.
35. Kolmogorov, A.N.; Fomin, S.V. *Introductory Real Analysis*; Dover Publication, Inc.: New York, NY, USA, 1970.
36. Alabdulmohsin, I.M. An information theoretic route from generalization in expectation to generalization in probability. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), Fort Lauderdale, FL, USA, 20–22 April 2017.
37. Alabdulmohsin, I. Information Theoretic Guarantees for Empirical Risk Minimization with Applications to Model Selection and Large-Scale Optimization. In Proceedings of the International Conference on Machine Learning (ICML 2018), Stockholm, Sweden, 10–15 July 2018; pp. 149–158.
38. Pavlovski, M.; Zhou, F.; Arsov, N.; Kocarev, L.; Obradovic, Z. Generalization-Aware Structured Regression towards Balancing Bias and Variance. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 2616–2622.
39. Alabdulmohsin, I.M. Algorithmic Stability and Uniform Generalization. In Proceedings of the NIPS 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 19–27.

40. Pelleg, D.; Moore, A.W. X-means: Extending k-means with efficient estimation of the number of clusters. In Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, USA, 29 June–2 July 2000; pp. 727–734.
41. Bardenet, R.; Maillard, O.A. Concentration inequalities for sampling without replacement. *Bernoulli* **2015**, *21*, 1361–1385. [CrossRef]