Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms

Milad Sefidgaran, Amin Gohari, Gael Richard, Umut Şimşekli

Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms

Milad Sefidgaran MILAD.SEFIDGARAN@TELECOM-PARIS.FR
LTCI, Télécom Paris, Institut Polyetechnique de Paris

Amin Gohari AMIN.AMINZADEH@GMAIL.COM
Tehran Institute for Advanced Studies, Khatam University

Gaël Richard GAEL.RICHARD@TELECOM-PARIS.FR
LTCI, Télécom Paris, Institut Polyetechnique de Paris

Umut Şimşekli UMUT.SIMSEKLI@INRIA.FR
INRIA & ENS – PSL Research University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Understanding generalization in modern machine learning settings has been one of the major challenges in statistical learning theory. In this context, recent years have witnessed the development of various generalization bounds suggesting different complexity notions such as the mutual information between the data sample and the algorithm output, compressibility of the hypothesis space, and the fractal dimension of the hypothesis space. While these bounds have illuminated the problem at hand from different angles, their suggested complexity notions might appear seemingly unrelated, thereby restricting their high-level impact. In this study, we prove novel generalization bounds through the lens of rate-distortion theory, and explicitly relate the concepts of mutual information, compressibility, and fractal dimensions in a single mathematical framework. Our approach consists of (i) defining a generalized notion of compressibility by using *source coding* concepts, and (ii) showing that the 'compression error rate' can be linked to the generalization error both in expectation and with high probability. We show that in the 'lossless compression' setting, we recover and improve existing mutual information-based bounds, whereas a 'lossy compression' scheme allows us to link generalization to the *rate-distortion dimension* – a particular notion of fractal dimension. Our results bring a more unified perspective on generalization and open up several future research directions.

Keywords: Generalization error, rate-distortion theory, source coding.

1. Introduction

Many important problems in statistical learning can be cast as the *population risk minimization* problem, which is defined as follows (Shalev-Shwartz and Ben-David, 2014):

$$\min_{w\in\mathcal{W}}\Big\{\mathcal{L}(w) \coloneqq \mathbb{E}_{Z\sim\mu}[\ell(Z,w)]\Big\}, \tag{1}$$

where $\mathcal{W}\subset\mathbb{R}^d$ denotes a parametric *hypothesis class*, $Z\in\mathcal{Z}$ denotes the *input data* with $\mathcal{Z}$ being the *data space*, μ denotes an unknown *data distribution* over $\mathcal{Z}$, and $\ell\colon \mathcal{Z}\times\mathcal{W}\to\mathbb{R}^+$ is a *loss*

function that measures the quality of a hypothesis $w \in \mathcal{W}$. As the data distribution μ is unknown in practice, we instead consider the *empirical risk minimization* problem, given as follows:

$$\min_{w\in\mathcal{W}}\Big\{\hat{\mathcal{L}}(S,w) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\ell(Z_i,w)\Big\}, \tag{2}$$

where $S := \{Z_1,\dots,Z_n\}$ denotes a *training dataset* with independent and identically distributed (i.i.d.) elements, i.e., each $Z_i \sim_{\text{i.i.d.}} \mu$.

To attack the optimization problem (2), arguably, the most common approach is to utilize a stochastic optimization algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ (e.g., stochastic gradient descent), such that the algorithm outputs a *random* hypothesis, i.e., $\mathcal{A}(S) = W \in \mathcal{W}$. One of the main challenges in statistical learning theory has been then to understand the behavior of the so-called *generalization error* associated with the algorithm output, that is the difference between the population and empirical risks induced by the algorithm output: $\operatorname{gen}(S,\mathcal{A}(S)) \coloneqq \mathcal{L}(\mathcal{A}(S)) - \hat{\mathcal{L}}(S,\mathcal{A}(S))$. It has been illustrated that classical algorithm-independent generalization bounds fall short at explaining the (perhaps unexpected) success of modern machine learning systems (Zhang et al., 2017). This has motivated the development of algorithm-dependent generalization bounds, a field that has been evolving in different directions.

An important direction in this context, and the one that is closest to our study, is based on analyzing the generalization error by using information-theoretic tools. Initiated by Russo and Zou (2016) and Xu and Raginsky (2017), these approaches link the generalization error to the *mutual information* between the data sample S and the algorithm output W; suggesting that a lower statistical dependence between S and W implies better generalization. Their initial results were later improved by using different conditional versions of the mutual information (Harutyunyan et al., 2021; Haghifam et al., 2021; Negrea et al., 2020b; Steinke and Zakynthinou, 2020; Bu et al., 2020; Haghifam et al., 2020), and were further generalized to more general notions of the mutual information that are defined through f-divergences (rather than the Kullback-Leibler divergence) (Esposito et al., 2020; Hellstrom and Durisi, 2020; Masiha et al., 2021).

A second approach has been based on the observation that the algorithm output W can be 'compressible' in different senses. Littlestone and Warmuth (1986) in a pioneer work, considered a compressibility framework for the binary classification problem, in which *compressed hypothesis* are chosen based on a subset of length k of S such that the picked hypothesis predicts correctly the label for all $Z_i \in S$. They showed that whenever such a compressing strategy exists, the algorithm generalizes well. The compressibility approach is later applied in different ways especially to over-parametrized neural networks (Arora et al., 2018; Suzuki et al., 2020a,b; Negrea et al., 2020a; Hsu et al., 2021; Barsbey et al., 2021; Baykal et al., 2019; Kuhn et al., 2021). Loosely speaking, under different compressibility assumptions for W, these studies showed that a higher level of compressibility indicates a lower generalization error since the hypothesis class $\mathcal{W}$ can be approximated by a smaller, 'compressed' space, which intuitively induces a lower worst-case error.

Finally, a recently initiated line of research has illustrated that when $\mathcal{A}$ is chosen as an iterative optimization algorithm, due to its recursive nature, $\mathcal{A}$ might generate a 'fractal structure', either in its optimization trajectories (Şimşekli et al., 2020; Birdal et al., 2021; Hodgkinson et al., 2021), or in the support of its stationary distribution (Camuto et al., 2021). These studies showed that the generalization error can be linked to the 'intrinsic dimension' of the fractal structure that is generated by the algorithm; suggesting that a smaller intrinsic dimension implies improved generalization.

Even though these three research directions have shed light on different façades of the problem of understanding the generalization error, the mathematical frameworks that underlie their theoretical results and their implied take-home messages might be seemingly unrelated, thereby restricting their high-level impact. In this paper, we prove novel generalization bounds through the lens of rate-distortion theory (Berger, 1975), and explicitly relate the concepts of mutual information, compressibility, and fractal dimensions in a single mathematical framework.

To achieve this goal, we first define a generalized notion of compressibility by using *source coding* concepts from information theory, which then allows us to use 'information-theoretic coverings' for $\mathcal{W}$ that we will detail in Section 3. Within this context, we show that the 'compression error rate' of an algorithm $\mathcal{A}$ can be linked to its generalization error both in expectation and with high probability. Next, we show that the aforementioned information-theoretical frameworks can be obtained as a special case of our setup, which is referred to as 'lossless compression'. Thanks to this connection, the results of Xu and Raginsky (2017) can be re-derived. The bound in (Xu and Raginsky, 2017, Theorem 1) is in terms of the *mutual information* between S and W, denoted as $I(S;W)$, which was previously viewed as the dataset dependency of the algorithm. However, our framework reveals that it is an upper-bound on the compression rate in terms of *lossless algorithm compressibility*. This new perspective allows us to introduce the notion of *lossy algorithm compressibility* to handle continuous or large alphabets where (Xu and Raginsky, 2017, Theorem 1) can be vacuous as $I(S;W)$ can be very large; implying that a large $I(S;W)$ does not necessarily indicate the algorithm will not generalize as long as the algorithm is 'lossily' compressible. We further established novel *tail bounds* suggesting that $I(S;W)$ (or its lossy version) needs to be small not only for the underlying distribution of (S,W), but also for any distribution in its vicinity. This is in the spirit of *stability*: the algorithm should be compressible under any small perturbation of the dataset and hypothesis. The new tail bounds are established using a new 'information-theoretic covering' technique, highlighted in Section 4.

Similarly, we derive and improve the results based on 'conditional mutual information' (Steinke and Zakynthinou, 2020) in Appendix A. Thanks to this approach, we established tail and in expectation bounds that recover the VC-dimension bounds (Corollaries 18 and 22); the recovery in terms of the tail bound is novel.

By exploiting the flexibility of our lossy compression framework, we further extend our results and obtain bounds in terms of the intrinsic dimension of the marginal distribution of W, namely the *rate-distortion dimension* (Kawabata and Dembo, 1994). Our results bring a unified perspective on mutual information, compressibility, and fractal dimensions, and open up several future research directions as we will point out in Section 5.

2. Preliminaries

2.1. Notation and problem setup

Random variables, their realizations, and their domains are denoted by upper-case letters, lower-case letters, and calligraphy fonts, *e.g.* X, x, and $\mathcal{X}$. We assume that all the domains are endowed with their Borel sigma fields. By P_X, we denote the distribution of X, defined on some measurable space $(\mathcal{X},\mathcal{F})$, and by $\operatorname{supp}(P_X)$ we denote its support. The expected value of X is denoted by $\mathbb{E}[X]$. We call a random variable X (absolutely) continuous if it admits a density with respect to the Lebesgue measure. The random variable X is called σ-subgaussian, if $\mathbb{E}[\exp(t(X-\mathbb{E}[X]))] \leqslant \exp(\sigma^2 t^2/2)$, $\forall t \in \mathbb{R}$. A collection of $m \in \mathbb{N}$ random variables is denoted by $X^m = (X_1,\ldots,X_m)$, or simply

by bold letters $\mathbf{X}$, when m is known by the context. A sequence of m real numbers $x_1, \dots, x_m$ is denoted by $\{x_i\}_{i=1}^m$. Similar conventions are used for sequences of sets or functions. The set of integers $\{1, \dots, m\}$ is denoted by $[m]$. We use $\mathbb{R}^+$ to denote nonnegative real numbers.

As mentioned in the introduction, we consider a generic *randomized algorithm* $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$, that has access to dataset $S = \{Z_1, \dots, Z_n\}$. This randomized algorithm induces a conditional distribution $P_{W|S}$. We denote the joint distribution of the dataset S and the hypothesis W by $P_{S,W} = \mu^{\otimes n} P_{W|S}$ and the marginal distribution of W by P_W.

Most of our results are expressed in terms of information-theoretic constructs, which we define as follows. For discrete random variables, the Shannon entropy function is defined as $H(X) \coloneqq \mathbb{E}[\log(1/P_X(X))]$. Similarly, conditional entropy is defined as $H(X|Y) \coloneqq \mathbb{E}[\log(1/P_{X|Y}(X|Y))]$. The mutual information between X and Y is defined as $I(X;Y) \coloneqq H(X) - H(X|Y)$, and intuitively measures the amount of *information* these random variables contain about each other. For continuous random variables, the differential entropy $h(X)$ is defined as $-D_{KL}(P_X \| \operatorname{Leb})$, where Leb is the Lebesgue measure on Euclidean spaces. In particular, if X has pdf p, then $h(X) = \mathbb{E}[\log(1/\mathrm{p}(X))]$, and $-\infty$ otherwise. Similarly, $h(X|Y)$ and $I(X;Y) \coloneqq h(X) - h(X|Y)$ are defined. The Kullback–Leibler (KL) divergence between two distributions Q and P defined on the same measurable space is defined as $D_{KL}(Q\|P) \coloneqq \mathbb{E}_Q\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P}\right]$, when $Q \ll P$, and equals ∞, otherwise. Here, $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is the Radon-Nikodym derivative of Q with respect to P.

2.2. Technical background on source coding

In this section, we will briefly review some results from the literature on source coding that will ease the introduction of our theoretical framework.[1] Consider a random variable W taking values in a finite set $\mathcal{W}$. It is well-known that one can represent W using $\lceil \log_2(|\mathcal{W}|) \rceil$[2] bits,[3] from which W can be recovered with no error. For instance, if W is a Bernoulli random variable, i.e., $\mathbb{P}(W = 1) = 1 - \mathbb{P}(W = 0) = \theta$, one bit suffices to represent W. However, intuitively speaking if θ is very close to zero or very close to one, using one full bit to represent W is wasteful because in such cases W is almost deterministic.

In his seminal paper, Shannon (1948) formalized this intuition by introducing the concept of *'block-coding'*, where he showed that the *'source'* W can be represented in a *compressed* way by using a significantly smaller number of bits, provided we can allow for a negligible probability of recovery error. The main idea behind block coding can be summarized as follows. As opposed to considering a single realization of the source W, we instead assume that we have access to a vector of m *independent* realizations of the source, denoted by W^m, and we are allowed to compress these m instances *simultaneously*. Moreover, the zero-error constraint in recovering m-instances is replaced by the 'asymptotically negligible error' criterion (i.e., the reconstruction error vanishes as $m \to \infty$). It turns out that joint description of such independent sources is more efficient than their individual description (Cover and Thomas, 2006). For instance, in our running example of W being a Bernoulli variable with parameter θ, W^m is a binary string of length m. By the law of large

1. For a more detailed introduction, we refer the reader to (Berger, 1975; Cover and Thomas, 2006; Csiszár and Körner, 2011; El Gamal and Kim, 2011; Polyanskiy and Wu, 2014).
2. For $a \in \mathbb{R}$, $\lceil a \rceil$ denotes the ceiling of a,*i.e.* $\min_{n \in \mathbb{N}}$, such that $n \geqslant a$.
3. Depending on the base of logarithm in the Shannon entropy function, the unit of information is either *bit* (base 2: $\log_2$) or *nat* (base e: $\log_e$). We state all results with base e for simplicity and compatibility with previous results. However, for the unit of information we use bit, as it is more common in the (digital) source coding context.

numbers, we expect W^m to have around $m\theta$ ones in it. So, even though there are 2^m binary strings of length m, roughly speaking W^m has about

$$\binom{m}{m\theta} \approx e^{mH(W)} \tag{3}$$

'effective' possibilities where in (3) we use Stirling's approximation of factorial to express the number of possibilities in terms of the $H(W)$, the Shannon entropy of W. Intuitively, for a discrete random variable W, 'the effective size' of m independent realizations is asymptotically about $e^{mH(W)}$, rather than $|\mathcal{W}|^m$. Concretely, there exist sets $\{\mathcal{C}_m\}_{m\in\mathbb{N}}, \mathcal{C}_m \subseteq \mathcal{W}^m$ with $|\mathcal{C}_m| = e^{m(H(W)+\varepsilon_m)}$ such that $\lim_{m\to\infty} \mathbb{P}(W^m \in \mathcal{C}_m) = 1$ and $\lim_{m\to\infty} \varepsilon_m = 0$. Thus, Shannon showed that the fundamental limit for the compression of information is determined by the Shannon entropy function $H(W)$, which can be much smaller than $\log(|\mathcal{W}|)$.

Unfortunately, the number of bits required to represent even a single realization of a continuous random variable W is infinity. For instance, if W is a uniform random variable on $[0, 1]$, we need infinitely many bits to convey it. However, one bit is enough to represent a single realization of W within distance 0.5 by mapping W to either $\hat{W} = 0$ if $W < 0.5$, and to $\hat{W} = 1$ if $W \geqslant 0.5$, and then conveying $\hat{W}$ instead of W. In this example, the reconstruction space (or the set of quantization points) is $\hat{\mathcal{W}} = \{0, 1\}$, the reconstruction is *lossy* (almost surely we never recover the original W) and we measure the distance (or the distortion) between W and its reconstruction $\hat{W}$ by $|W - \hat{W}|$.

Let us begin by describing the lossy compression of a *single instance* of an arbitrary random variable W. In many information-theoretic and signal processing applications, it suffices to recover a *distorted* version $\hat{W} \in \hat{\mathcal{W}}$ of W, as long as the incurred distortion is within an 'acceptable' range, *i.e.* for $\epsilon \geqslant 0$ and a chosen *distortion function* $\varrho\colon \mathcal{W} \times \hat{\mathcal{W}} \to \mathbb{R}^+$, we have $\varrho(W; \hat{W}) \leqslant \epsilon$. While the quantized (distorted) space $\hat{\mathcal{W}}$ is equal to $\mathcal{W}$ in many cases, it can be different in general.[4] To facilitate the explanation, assume $\hat{\mathcal{W}} = \mathcal{W}$ for the rest of this section, *i.e.,* we are required to produce $\hat{W} \in \mathcal{W}$. Next, let us consider the case of block coding where instead of a single realization of the source W, we have access to W^m which is a vector of m *independent* realizations of the source. This problem is known as the vector-quantization problem. Intuitively speaking, to compress W^m we can take a collection of k quantization points $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_k$ in $\mathcal{W}^m$, where $\hat{\mathbf{w}}_j = (\hat{w}_{j,1}, \ldots, \hat{w}_{j,m})$ for $j \in [k]$, and map $W^m = (W_1, \ldots, W_m)$ to its closest quantization point. Here, the distortion function between W^m and a quantization point $\hat{\mathbf{w}}_j$ should be defined; it is often chosen to be the average of coordinate-wise distortions:

$$\frac{1}{m}\sum_{i=1}^{m} \varrho(W_i, \hat{w}_{j,i}).$$

Because the number of quantization points is k, we require $\log_2(k)$ bits to convey the index of the quantization point. The ratio $\log_2(k)/m$ is called the compression *rate*, because it represents the number of compression bits per source realization. For the selection of quantization points $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_k$ to succeed, we can consider balls of radius ϵ around these quantization points and require that with high probability W^m falls into the union of these balls. This can be seen as a "block covering" of $\mathcal{W}^m$ with average distortion ϵ. Note that block covering may need a smaller number of quantization points than the case where the complete covering of the space $\mathcal{W}^m$ with the worst-case distortion ϵ is required, as in ϵ-net coverings (Anthony and Bartlett, 1999).

4. For example, to convey the sign of $W \in \mathcal{W} = \mathbb{R}$, it is natural to consider $\hat{\mathcal{W}} = \{-1, +1\}$ and $\varrho(w, \hat{w}) = \mathbb{1}_{\{w\hat{w}<0\}}$.

In this context, the goal becomes finding the minimum number of bits that is required to compress i.i.d. repetitions of the source W so that it can be recovered within a given distortion margin. Shannon (1948) showed that the minimum compression rate needed for recovering a source with distortion ϵ is determined by the *rate-distortion function*

$$\mathfrak{RD}(\epsilon; P_W, \varrho) \coloneqq \inf I(W;\hat{W}), \quad \text{such that} \quad \mathbb{E}_{W,\hat{W}}[\varrho(W,\hat{W})] \leqslant \epsilon,$$

where the infimum is over all conditional probability distributions (Markov kernels) $P_{\hat{W}|W}$. Specifically, Shannon showed that for any *rate* $R > \mathfrak{RD}(\epsilon; P_W, \varrho)$, a sequence of *quantization codebooks* $\{\mathcal{C}_m\}_{m\in\mathbb{N}}$, $\mathcal{C}_m{=}\{\hat{\mathbf{w}}_j{=}(\hat{w}_{j,1},\dots,\hat{w}_{j,m}), j \in [l_m]\} \subseteq \mathcal{W}^m$ exists such that $l_m \leqslant e^{mR}$ and

$$\lim_{m\to\infty} \mathbb{P}_{W^{\otimes m}}\left(\forall j\colon \frac{1}{m}\sum\nolimits_{i=1}^{m} \varrho(W_i,\hat{w}_{j,i}) > \epsilon\right) = 0. \tag{4}$$

Intuitively, for discrete variables, the *effective size* of m independent realizations of W is about $e^{mH(W)}$, and each codeword $\hat{\mathbf{w}}_j$ covers about $e^{mH(W|\hat{W})}$ of them, and thus, the total needed codewords to cover $\mathcal{W}^m$ with high probability is about $e^{mI(W;\hat{W})}$. Similar intuition holds for the continuous W, by considering $h(W)$ and $h(W|\hat{W})$, and by considering the *effective volume* of W^m.

Finally, a series of works, *e.g.* (Marton, 1974; Han, 2000; Iriyama, 2005; Bakshi and Bansal, 2005), studied the rate of convergence of the probability in (4) to zero for a fixed rate R. Equivalently, one can formulate this problem as the minimum needed rate R to have the above error probability decaying at least as fast as δ^m. For sources with finite alphabets, this quantity is equal to (Marton, 1974, Theorem 1) $\sup_{Q\colon D_{KL}(Q\|P_W)\leqslant\log(1/\delta)} \mathfrak{RD}(\epsilon; Q, \varrho)$, where the supremum is over all distributions Q defined over $\mathcal{W}$ such that $Q \ll P_W$. Intuitively, the empirical distribution $\hat{P}_{W^m}$ of a vector of realizations W^m satisfies $D_{KL}(\hat{P}_{W^m}\|P_W) \leqslant \log(1/\delta)$, with probability at least $1-\delta^m$. The idea is to "cover" all such high probable realizations in the balls with radius ϵ. It turns out the needed rate is the supremum of the needed rate for each empirical distribution Q. Similar error exponent term for continuous sources can be found in (Iriyama, 2005, Theorem 1).

In this work, we apply source coding concepts and techniques to establish bounds on the generalization error. To this end, we attempt to 'reliably compress' the hypothesis space with respect to a distortion that depends on the excess generalization error induced by compression. We allow the compression to be lossy within a distortion level. Then, we establish bounds on the generalization error in terms of the compression rate, amount of distortion, and reliability level.[5]

3. Generalization Bounds via Rate Distortion Theory

We start by explaining our notion of compressibility adapted to algorithms.

3.1. Compressibility of an algorithm

The compression, in its classical source coding sense, aims to save a compressed version of a source that is *close enough* to the source and requires a smaller storage capacity. Similarly, for a learning

5. The rate-distortion theory was previously used in (Bu et al., 2021; Masiha et al., 2021). For instance, in (Bu et al., 2021), it is used to compare the expectation of the generalization error of a *compressed* learning model with respect to the *original* model. Herein, we use it to analyze the generalization performance of the *original* learning model. In Masiha et al. (2021), generalization error is related to the rate-distortion theory by noting the similarity of the related formulas. The connection provided in this work is operational and thus much deeper.

algorithm $\mathcal{A}\colon \mathcal{S} \to \mathcal{W}$, where $\mathcal{S} = \mathcal{Z}^n$, by having a dataset S and a picked hypothesis choice $\mathcal{A}(S) = W \in \mathcal{W}$, we are interested in finding another algorithm $\hat{\mathcal{A}}(S, W) = \hat{W} \in \hat{\mathcal{W}} \subseteq \mathcal{W}$ that has fewer number of probable output hypotheses and *performs closely* to the original algorithm. In this work, we consider the generalization error as the compression performance. Consider a training dataset s and two hypotheses w and $\hat{w}$. We define the distortion function $d\colon \mathcal{W} \times \hat{\mathcal{W}} \times \mathcal{S} \to \mathbb{R}$ between these two pairs of realizations as the difference of their generalization performances:[6]

$$d(w, \hat{w}; s) \coloneqq \operatorname{gen}(s, w) - \operatorname{gen}(s, \hat{w}). \tag{5}$$

Note that here, unlike the source-coding literature, we allow the distortion function to take negative values.

To guarantee that this distortion (between single outputs of the original and compressed algorithms) does not exceed a threshold, we need to control the *worst-case* distortion caused by compression, among all probable w and $\hat{w}$, which might end up with overly pessimistic results. To avoid this, we utilize the *block coding* technique as follows. For a block of $m \in \mathbb{N}$ independent datasets $s^m = (s_1, \ldots, s_m)$ and a block of picked hypotheses $W^m = (W_1, \ldots, W_m)$, where W_i is a hypothesis choice based on dataset s_i, i.e. $\mathcal{A}(s_i) = W_i$, $i \in [m]$, with a slight abuse of notations, denote $\mathcal{A}(s^m) = W^m$. We then consider a compression algorithm $\hat{\mathcal{A}}_m\colon \mathcal{S}^m \times \mathcal{W}^m \to \hat{\mathcal{W}}^m$ that takes as input particular realizations (s^m, w^m), where $\mathcal{A}(s^m) = w^m$, and outputs a block of hypotheses $\hat{W}^m = (\hat{W}_1, \ldots, \hat{W}_m)$. We also need to extend our definition in (5) to measure the distortion between two blocks of algorithm realizations $\mathcal{A}(s^m) = w^m$ and $\hat{\mathcal{A}}_m(s^m, w^m) = \hat{w}^m$. For now, let us use $d_m\colon \mathcal{S}^m \times \mathcal{W}^m \times \hat{\mathcal{W}}^m \to \mathbb{R}$ for $m \in \mathbb{N}$ to denote this extended distortion function, whose details will be provided in the next section. In particular, we will use the extended distortion function defined in (8) to obtain in expectation bounds and an alternative definition given in (12) to obtain tail bounds on the generalization gap.

Next, we define our compression algorithm. Fix a set $\mathcal{H}_m \subseteq \hat{\mathcal{W}}^m$, that we coin a *hypothesis book* (as an analogy to code book), and denote its cardinality by l_m. Denote the elements of $\mathcal{H}_m$ by $\hat{\mathbf{w}}_j = (\hat{w}_{j,1}, \ldots, \hat{w}_{j,m}) \in \hat{\mathcal{W}}^m$, where $j \in [l_m]$, i.e., $\mathcal{H}_m = \{\hat{\mathbf{w}}_j\}_{j=1}^{l_m}$. Having defined a distortion function d_m and fixed a set $\mathcal{H}_m$, among all compression algorithms $\hat{\mathcal{A}}_m$ such that $\hat{\mathcal{A}}_m(s^m, w^m) \in \mathcal{H}_m$, we consider the *optimal* compression algorithm, denoted by $\hat{\mathcal{A}}_m^*(s^m, w^m; \mathcal{H}_m) = \hat{\mathbf{w}}_j$, where $j = \arg\min_{j \in [l_m]} d_m(w^m, \hat{\mathbf{w}}_j; s^m)$. With this choice and for a fixed distortion level ϵ, we define the error event that happens when the average distortion between the original and the optimal compressed algorithm exceeds ϵ:

$$\mathcal{E}_m(\mathcal{H}_m, \epsilon; d_m) \coloneqq \left\{\min_{j \in [l_m]} d_m(\mathcal{A}(S^m), \hat{\mathbf{w}}_j; S^m) > \epsilon\right\}. \tag{6}$$

Now, we are ready to define our compressibility notion, which will lay the basis of our generalization bounds.

Definition 1 *The learning algorithm $\mathcal{A}$ is $(R, \epsilon; \{d_m\}_m)$-compressible[7] for some $R \in \mathbb{R}^+$ and $\epsilon \in \mathbb{R}$, if there exists a sequence of hypothesis books $\{\mathcal{H}_m\}_{m \in \mathbb{N}}$, $\mathcal{H}_m = \{\hat{\mathbf{w}}_j, j \in [l_m]\} \subseteq \hat{\mathcal{W}}^m$ such that $l_m \leqslant e^{mR}$ and*

$$\lim_{m \to \infty} \mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m(\mathcal{H}_m, \epsilon; d_m)) = 0, \tag{7}$$

6. While d is clearly not a metric, it also depends on the underlying distribution μ; we drop this dependence for ease of notations.
7. While many terms in this work, including R and the rate-distortion terms in the rest of the text, depend on μ, $P_{W|S}$, n, and the loss function, we drop these dependencies for ease of exposition.

where $\mathcal{E}_m(\mathcal{H}_m, \epsilon; d_m)$ is defined in (6) and $\mathbb{P}_{(S,W)^{\otimes m}}$ denotes the m-times product measure of the joint distribution of W and S.

3.2. Bounds on the expected value of the generalization gap

In this section, we prove bounds on the expected generalization error, provided $\mathcal{A}$ is compressible. Intuitively, we first find compression schemes that *cover* (S^m, W^m) with high probability, in a sense that is defined in (7), such that on average the difference of generalization errors of the original and compressed algorithms does not exceed a threshold. Then, we show that the expected generalization error can be bounded in terms of the parameters of this compressed algorithm. To do so, by borrowing from the source coding literature, we define a distortion function between m realizations of the two algorithms as:

$$\vartheta_m(w^m, \hat{w}^m; s^m) := \frac{1}{m} \sum_{i=1}^{m} d(w_i, \hat{w}_i; s_i), \tag{8}$$

where $d(w, \hat{w}; s)$ was defined in (5).

Having condition (7) for this distortion function guarantees that the expectation of the difference of the generalization errors of the original and compressed algorithms does not exceed ϵ. This is stated in Lemma 27, which is used in the proof of the following result, proved in Appendix E.1.

Theorem 2 *If a learning algorithm $\mathcal{A}(S)$ is $(R, \epsilon; \{|\vartheta_m|\}_m)$-compressible,[8] if $\mathbb{E}_{S,W}[|\operatorname{gen}(S,W)|] < \infty$, and if for all $w \in \mathcal{W}$, $\ell(Z, w)$ is σ-subgaussian, then $|\mathbb{E}[\operatorname{gen}(S,W)]| \leqslant \sqrt{2\sigma^2 R/n} + \epsilon$.*

This result shows that the compressibility of an algorithm directly translates into having a good generalization performance, which can be seen as an information theoretic counterpart of the existing compression bounds, e.g., (Arora et al., 2018; Suzuki et al., 2020b). To make the above bound more explicit, we establish the following bound on the compressibility of any arbitrary algorithm, whose proof is given in Appendix E.2. Let[9]

$$R_E(\epsilon) = \inf_{P_{\hat{W}|S}} I(S; \hat{W}), \quad \text{such that} \quad \left|\mathbb{E}\Big[\operatorname{gen}(S,W) - \operatorname{gen}(S,\hat{W})\Big]\right| \leqslant \epsilon, \tag{9}$$

where the expectation is with respect to $P_{S,W}$ and $P_S \times P_{\hat{W}|S}$.

Theorem 3 *Assume that the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$, where $\mathcal{S}$ and $\mathcal{W}$ are finite sets. Then, for every $\epsilon \in \mathbb{R}$ and any $\nu_1, \nu_2 > 0$, the algorithm $\mathcal{A}(s)$ is $(R_E(\epsilon) + \nu_1, \epsilon + \nu_2; \{|\vartheta_m|\}_m)$-compressible.*

This theorem can be extended to infinite sets, with some further assumptions on separability of $\operatorname{gen}(S,W)$ with respect to (S,W) and using the quantization technique used in the proof of (El Gamal and Kim, 2011, Theorem 3.6). Now, combining Theorems 2 and 3 yields:

Theorem 4 *Assume that the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and for all $w \in \mathcal{W}$, $\ell(Z, w)$ is σ-subgaussian. Then, for any $\epsilon \in \mathbb{R}$, $|\mathbb{E}[\operatorname{gen}(S,W)]| \leqslant \sqrt{2\sigma^2 R_E(\epsilon)/n} + \epsilon$.*

The extended versions of Theorems 2 and 4 that include bounds on $\mathbb{E}[\operatorname{gen}(S,W)]$ and $\mathbb{E}[|\operatorname{gen}(S,W)|]$ as well, can be found in Appendix D.1.[10]

8. By $|\vartheta_m|$ we mean simply the distortion function which is equal to $|\vartheta_m(w^m, \hat{w}^m; s^m)|$ for any $w^m, \hat{w}^m, s^m$.
9. Intuitively, $\mathbb{E}\Big[\operatorname{gen}(S,W) - \operatorname{gen}(S,\hat{W})\Big]$ can be seen as the limit of the distortion function ϑ_m when $m \to \infty$.
10. The mild sufficient condition $\mathbb{E}[|\operatorname{gen}(S,W)|] < \infty$ is used to bound $\mathbb{E}[|\operatorname{gen}(S,W)|]$, $\mathbb{E}[\operatorname{gen}(S,W)]$, and $|\mathbb{E}[\operatorname{gen}(S,W)]|$ in the extended version of Theorem 2. The sufficiency of the condition $|\mathbb{E}[\operatorname{gen}(S,W)]| < \infty$ for bounding $|\mathbb{E}[\operatorname{gen}(S,W)]|$, although seemingly true, is not shown in this work.

In addition to finite sets, Theorem 4 can be derived using Theorems 2 and 3 also for the infinite sets that satisfy some further separability assumptions. However, for the infinite set, without any further assumptions, we show this alternatively and trivially in Appendix E.3 by using and extending the existing results of Xu and Raginsky (2017, Theorems 1, 4), corresponding to Theorem 4 with $\epsilon = 0$ (see Corollary 5 in below). Theorem 4 is extended similarly to (Bu et al., 2020) in Theorem 25 (Appendix D.1) that recovers and potentially improves over (Bu et al., 2020, Proposition 1).

Corollary 5 *Suppose the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and the loss function $\ell(Z, w)$ is σ-subgaussian for any $w \in \mathcal{W}$. Then, $|\mathbb{E}[\operatorname{gen}(S, W)]| \leqslant \sqrt{2\sigma^2 I(S;W)/n}$.*

In this corollary, by applying a compressibility approach we could recover the results obtained using the Donsker–Varadhan's identity. Indeed, in Appendix B.1 we showed that this identity can be interpreted and derived via a compressibility approach.

The case of $\epsilon = 0$, considered in (Xu and Raginsky, 2017; Bu et al., 2020), corresponds to the lossless compression in source coding. While for countable sets of S or W, we can reliably cover (S^m, W^m) with $\epsilon = 0$ (in the sense of (7)) and bounded R, for continuous sources and hypotheses, this term could be infinite. In contrast, considering $\epsilon \neq 0$, corresponds to the lossy compression in source coding, which allows to reliably cover (S^m, W^m) with bounded R within distortion ϵ. Note that even for countable sets, $\epsilon \neq 0$ can give better bounds.[11]

The benefit of $\epsilon \neq 0$ becomes more clear by having a Lipschitz loss assumption. Combining this assumption with the above theorem directly yields an upper bound on the expected generalization error in terms of the rate-distortion function of the hypothesis.

Corollary 6 (Lipschitz loss) *Suppose that for a distortion function $\varrho \colon \mathcal{W} \times \hat{\mathcal{W}} \to \mathbb{R}^+$ and every $z, w, \hat{w}$, $|\ell(z, w) - \ell(z, \hat{w})| \leqslant \mathfrak{L}\varrho(w, \hat{w})$[12] and $\ell(Z, w)$ is σ-subgaussian. Then, for any $\epsilon \in \mathbb{R}^+$, we have $|\mathbb{E}[\operatorname{gen}(S, W)]| \leqslant \sqrt{2\sigma^2 \mathfrak{RD}(\epsilon/(2\mathfrak{L}); P_W, \varrho)/n} + \epsilon$, where $\mathfrak{RD}(\epsilon; P_W, \varrho)$ is the rate-distortion function with respect to the distortion function ϱ:*

$$\mathfrak{RD}(\epsilon; P_W, \varrho) \coloneqq \inf_{P_{\hat{W}|W}} I(W; \hat{W}), \quad \text{such that} \quad \mathbb{E}_{W,\hat{W}}\Big[\varrho(W, \hat{W})\Big] \leqslant \epsilon. \tag{10}$$

While the term $R_E(\epsilon)$ in (9) is in general intractable, the above bound is amenable to computation once the marginal distribution P_W is known; a more relaxed constraint than knowing $P_{S,W}$ which is needed in many of the information-theoretic bounds on generalization error. The rate-distortion computation is a convex minimization problem over $P_{\hat{W}|W}$ and ϵ, that can be effectively computed for finite alphabets using Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972). Note that using Carathéodory's theorem (El Gamal and Kim, 2011, Appendix C), it can be shown that it is sufficient to consider $\hat{\mathcal{W}}$ such that $|\hat{\mathcal{W}}| \leqslant |\mathcal{W}| + 1$. For the continuous alphabets, this terms can be efficiently estimated using the fine quantization technique (*e.g.* (El Gamal and Kim, 2011, Proof

11. The approach applied in in (Negrea et al., 2020a) for studying $\mathbb{E}[\operatorname{gen}(S, W)]$ also can be seen as lossy compression. They considered choosing a randomized 'surrogate hypothesis' $\hat{W}$ for each W, and argue that to establish a good bound on $\mathbb{E}[\operatorname{gen}(S, W)]$, one could benefit from the trade-off between $\mathbb{E}[\hat{\mathcal{L}}(S, \hat{W}) - \hat{\mathcal{L}}(S, W)]$ and $\mathbb{E}[\mathcal{L}(\hat{W}) - \hat{\mathcal{L}}(S, \hat{W})]$. However, they have not proposed general explicit bounds on these terms, and rather considered ad-hoc strategies for overparameterized linear regression and hypercube classification problems, when $\hat{\mathcal{L}}(S, W) = 0$.

12. Note that this condition and imposing the Markov chain $\hat{W} - W - S$ yield $\Big|\mathbb{E}\Big[\operatorname{gen}(S, W) - \operatorname{gen}(S, \hat{W})\Big]\Big| \leqslant 2\mathfrak{L}\mathbb{E}_{W,\hat{W}}\Big[\varrho(W, \hat{W})\Big]$ and $I(S; \hat{W}) \leqslant I(W; \hat{W})$ by data processing inequality.

of Theorem 3.6) or by using the existing lower bounds, *e.g.* (Riegler et al., 2018), that are almost tight in the small ϵ regime.

As an analytical example, suppose that the data $Z \in \mathbb{R}^d$ is composed of d i.i.d. elements, each one distributed according to the normal distribution $\mathcal{N}(0, \sigma_0^2)$ and suppose that we choose W as $W = \frac{1}{n}\sum_{i=1}^n Z_i \sim \mathcal{N}(0, (\sigma_0^2/n)\mathrm{I}_d)$, where I_d is the $d \times d$ identity matrix. Further, suppose that $\ell(Z, w)$ is σ-subgaussian for any $w \in \mathbb{R}^d$ and $|\ell(z, w) - \ell(z, \hat{w})| \leqslant \mathfrak{L}\|w - \hat{w}\|^2$. Then, while $I(S; W) = \infty$, Corollary 6 together with (Cover and Thomas, 2006, Theorem 10.3.2) yield $|\mathbb{E}[\mathrm{gen}(S, W)]|$ is bounded by $\min_{0<\epsilon\leqslant d\sigma_0^2/n} \sqrt{\sigma^2 d \log\big(2\mathfrak{L}d\sigma_0^2/(n\epsilon)\big)/n + \epsilon}$, which equals $2\mathfrak{L}d\sigma_0^2/n$ for $\epsilon = 2\mathfrak{L}d\sigma_0^2/n$.

The optimal order of ϵ (and the corresponding rate-distortion terms) with respect to n depends on $P_{S,W}$, as well as the loss function. In the above example, ϵ is chosen as $\mathcal{O}(1/n)$, and in the following corollary as $\mathcal{O}(1/\sqrt{n})$, resulting in the rate-distortion terms of order $\mathcal{O}(1)$ and $\mathcal{O}(\log(n))$, respectively.

In our next result, we show that our bound in terms of the rate-distortion function yields a fractal dimension-based bound as well. Let us define the rate-distortion dimension (Kawabata and Dembo, 1994) for a distribution Q as $\dim_{\mathrm{R}}(Q) \coloneqq \limsup_{\epsilon\to 0} \mathfrak{RD}(\epsilon; Q, \varrho)/\log(1/\epsilon)$.

Corollary 7 (Rate-distortion dimension) *Suppose that for a distortion function $\varrho\colon \mathcal{W}\times\hat{\mathcal{W}} \to \mathbb{R}^+$ and every $z, w, \hat{w}$, $|\ell(z, w) - \ell(z, \hat{w})| \leqslant \mathfrak{L}\varrho(w, \hat{w})$ and $\ell(Z, w)$ is σ-subgaussian. Moreover, assume that $\sup_{\epsilon\leqslant\epsilon_0} \mathfrak{RD}(\epsilon; P_W, \varrho)/\log(1/\epsilon)$ converges uniformly over n as $\epsilon_0 \to 0$. Then, there exists a n_0 such that for every $n \geqslant n_0$, $|\mathbb{E}[\mathrm{gen}(S, W)]| \leqslant \sqrt{4\sigma^2 \dim_{\mathrm{R}}(P_W)\log(n\mathfrak{L}^2)/n}$.*

This corollary, proved in Appendix E.4, shows the relation of our approach with dimension-based bounds. The rate-distortion dimension is a lower bound to the Minkowski (box-counting) dimension of the set $\mathcal{W}$ (Kawabata and Dembo (1994)), which was considered in (Şimşekli et al., 2020; Birdal et al., 2021). Moreover $\dim_{\mathrm{R}}(P_W)$ is equal to the Rényi information dimension under certain conditions (Kawabata and Dembo, 1994, Proposition 3.3). The latter dimension is shown to be related to the fundamental limits of the almost lossless compression (Wu and Verdú, 2010).

3.3. Tail bounds on the generalization gap

To establish a tail bound on the generalization performance, we need to find a compression scheme that not only *covers* (S^m, W^m) with high probability (in a sense of (7)), but also its probability of covering failure is exponentially decreasing with m, which leads us to the following notion.

Definition 8 *The learning algorithm $\mathcal{A}$ is $(R, \epsilon, \delta; \{d_m\}_m)$- exponentially compressible for some $\delta > 0$, if conditions of Definition 1 hold and*

$$\lim_{m\to\infty}\left[-\frac{1}{m}\log\big(\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m(\mathcal{H}_m, \epsilon; d_m))\big)\right] \geqslant \log(1/\delta). \tag{11}$$

In other words, the error probability is asymptotically bounded by δ^m.

On the other hand, instead of considering (8), it turns out that it is sufficient to keep the difference between the *average* generalization error of the compressed algorithm and the *lowest* error of the

original algorithm within a threshold. More precisely, we define the new distortion function:[13]

$$\varphi_m(w^m, \hat{w}^m; s^m) \coloneqq \min_{j\in[m]} \operatorname{gen}(s_j, w_j) - \frac{1}{m}\sum_{i=1}^{m} \operatorname{gen}(s_i, \hat{w}_i). \tag{12}$$

By using this notion, our first tail bound on the generalization performance of an algorithm is stated in the following theorem, which is proved in Appendix E.5.

Theorem 9 *If a learning algorithm $\mathcal{A}$ is $(R, \epsilon, \delta; \{\varphi_m\}_m)$-exponentially compressible and if for all $w \in \mathcal{W}$, the loss $\ell(Z, w)$ is σ-subgaussian, then with probability at least $1-\delta$, we have that $\operatorname{gen}(S, W) \leqslant \sqrt{2\sigma^2(R + \log(1/\delta))/n} + \epsilon$.*

This result shows that exponentially compressible algorithms, with small (R, ϵ), generalize well with probability $1-\delta$. Next, in our main tail bound, we will show that any arbitrary algorithm is exponentially compressible, and we will establish a bound on its compressibility triplet (R, ϵ, δ). To state this result, we need some definitions. For a given distribution Q over $\mathcal{S} \times \mathcal{W}$, let

$$d_Q(\hat{w}; s) \coloneqq \inf_{(s', w')\in \operatorname{supp}(Q)} \big[\operatorname{gen}(s', w')\big] - \operatorname{gen}(s, \hat{w}). \tag{13}$$

Intuitively, $\mathbb{E}_{S,\hat{W}}[d_Q(\hat{W}; S)]$ can be seen as the limit of the distortion function φ_m when $m \to \infty$. Moreover, for a distribution Q defined over $\mathcal{S} \times \mathcal{W}$, let

$$\mathfrak{RD}^*(\epsilon; Q) \coloneqq \inf_{\substack{P_{\hat{W}|S}:\\ \mathbb{E}\left[d_Q(\hat{W};S)\right]\leqslant\epsilon}} I(S; \hat{W}) \leqslant \inf_{\substack{P_{\hat{W}|S}:\\ \mathbb{E}\left[\operatorname{gen}(S,W)-\operatorname{gen}(S,\hat{W})\right]\leqslant\epsilon}} I(S; \hat{W}), \tag{14}$$

where the infimum is over all Markov kernels $P_{\hat{W}|S}\colon \hat{\mathcal{W}} \times \mathcal{S} \to \mathbb{R}^+$ and the expectations and the mutual information are with respect to joint distributions Q and $Q_S \times P_{\hat{W}|S}$, where Q_S is the marginal distribution of S. Now, we state our main tail bound result, proved in Appendix E.6.

Theorem 10 *Suppose that the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and for all $w \in \mathcal{W}$, $\ell(Z, w)$ is σ-subgaussian. Then, for every $\epsilon \in \mathbb{R}$ and $\delta \geqslant 0$, with probability at least $1-\delta$,*

$$\operatorname{gen}(S, W) \leqslant \sqrt{\frac{2\sigma^2(R_p(\delta, \epsilon) + \log(1/\delta))}{n}} + \epsilon, \; R_p(\delta, \epsilon) \coloneqq \sup_{Q\colon D_{KL}(Q\|P_{S,W})\leqslant\log(1/\delta)} \mathfrak{RD}^*(\epsilon; Q),$$

where the supremum is over all probability distributions Q over $\mathcal{S} \times \mathcal{W}$.

To the best of our knowledge, this is the first information-theoretic tail bound on the generalization error with the logarithmic dependence on $1/\delta$. The bound does not reduce to previous results even for $\epsilon = 0$. In this case, $\mathfrak{RD}^*(0; Q) \leqslant I_Q(S; W)$ as $\hat{\mathcal{W}} = \mathcal{W}$ and $P_{\hat{W}|S} = Q_{W|S}$ are valid choices, where $I_Q(S; W)$ implies the mutual information under the distribution Q. Hence, as a corollary of the above theorem, with probability at least $1-\delta$, we have

$$\operatorname{gen}(S, W) \leqslant \sqrt{2\sigma^2\Big(\sup_{Q\colon D_{KL}(Q\|P_{S,W})\leqslant\log(1/\delta)} I_Q(S; W) + \log(1/\delta)\Big)/n}. \tag{15}$$

13. Note that $\varphi_m(w^m, \hat{w}^m; s^m) \leqslant \vartheta_m(w^m, \hat{w}^m; s^m)$. For further discussion on this distortion function, refer to Section 4.

The bound in Theorem 10 does not only depend on $P_{S,W}$, but on all Q close to $P_{S,W}$. This is similar to the error exponent result of (Marton, 1974, Theorem 1). Intuitively, by considering all Q satisfying $D_{KL}(Q\|P_{S,W}) \leqslant \log(1/\delta)$, we cover realizations of (S^m, W^m) with probability at least $1-\delta^m$. In other words, to have a good generalization bound with high probability, the algorithm should be compressible under all such Q that are close enough to $P_{S,W}$.

Theorem 10 does not take into account any additional stochasticity of the algorithm, as considered in (Harutyunyan et al., 2021). Considering such a scenario yields stronger results, presented in Appendix D.2.

Similar to the in expectation part, by having a Lipschitzness property and using (14), Theorem 10 can be upper-bounded in terms of the rate-distortion functions of the hypothesis set.

Corollary 11 (Lipschitz loss) *Suppose that for a distortion function $\varrho\colon \mathcal{W}\times\hat{\mathcal{W}}\to\mathbb{R}^+$ and every $z, w, \hat{w}$, $|\ell(z,w)-\ell(z,\hat{w})| \leqslant \mathfrak{L}\varrho(w,\hat{w})$ and $\ell(Z,w)$ is σ-subgaussian. Then, for any $\epsilon\in\mathbb{R}^+$, the term $R_p(\delta,\epsilon)$ in Theorem 10 can be upper bounded by*

$$R_p(\delta,\epsilon) \leqslant \sup_{Q_W\colon\, D_{KL}(Q_W\|P_W)\leqslant\log(1/\delta)} \mathfrak{RD}(\epsilon/(2\mathfrak{L}); Q_W, \varrho), \tag{16}$$

where the supremum is over all possible distributions Q_W over $\mathcal{W}$.

The above corollary recovers some classical results, *e.g.* the bound obtained by using ϵ-net coverings. This result, together with some other concrete examples are presented in Appendix D.3. Furthermore, similar to Corollary 7, one can derive a dimension-based bound by using Corollary 26.

Note that Theorem 10 can be made data-dependent using ideas of Negrea et al. (2020b). Finally, we further extend our results in Appendix A, that recovers (and improves in specific cases) the conditional mutual information based results of Steinke and Zakynthinou (2020); Harutyunyan et al. (2021).

4. Proof Outline

Our main results are new bounds on the generalization gap. However, we also develop new techniques which are rather general and applicable to arbitrary random variables. In particular, we derive a variational representation of the tail probability in Lemma 24 (Appendix C) that results the following tail bound for any arbitrary random variable X:

Theorem 12 *For arbitrary random variables $X\in\mathbb{R}$, $Y\in\mathcal{Y}$ distributed according to $(X,Y)\sim\mu_{X,Y}$, with marginals $X\sim\mu_X$ and $Y\sim\mu_Y$ and for any $\delta\geqslant 0$ and $\epsilon,\Delta\in\mathbb{R}$, we have*

$$\log\mu_X([\Delta,\infty)) \leqslant \tag{17}$$
$$\max\left[\log(\delta), \sup_{\nu_{X,Y}\in\mathcal{G}} \inf_{p_{\hat{X}|Y}\in\mathcal{Q}(\nu)} \inf_{q_{\hat{X}|Y},\,\lambda\geqslant 0}\left\{D_{KL}(p_{\hat{X}|Y}\nu_Y\|q_{\hat{X}|Y}\nu_Y) - \lambda(\Delta-\epsilon) + \log\mathbb{E}_{\mu_Y q_{\hat{X}|Y}}[e^{\lambda\hat{X}}]\right\}\right]$$

where $\mathcal{G} \coloneqq \{\nu_{X,Y} : D_{KL}(\nu_{X,Y}\|\mu_{X,Y}) \leqslant \log(1/\delta)\}$, $\hat{X}$ is a real valued random variable and $\mathcal{Q}(\nu)$ is the set of all conditional distributions $p_{\hat{X}|Y}$ such that under the joint distribution $p_{\hat{X}|Y}\nu_{X,Y}$ we have: $\left[\inf_{x\in\mathrm{supp}(\nu_X)} x\right] - \mathbb{E}[\hat{X}] \leqslant \epsilon$, ν_X and ν_Y are marginals of $\nu_{X,Y}$ with respect to X and Y, respectively, and the inner infimum is over all conditional distributions $q_{\hat{X}|Y}$.

This theorem is proved in Appendix E.7 and implies Theorem 10 by considering X as $\mathrm{gen}(S, W)$. A key idea used in this paper is leveraging the block covering technique to establish tail bounds. In the following subsection, we explain our general approach for this.

4.1. Tail bound via information-theoretic covering

Covering is a technique that allows to provide upper bounds on the tail probability or the expectation of an arbitrary random variable. The standard covering technique works as follows (see (Vershynin, 2018, Chapter 7)): consider a random process $(Y_t)_{t\in T}$, and the random variable $X = \sup_{t\in T} Y_t$. An ϵ-covering (or ϵ-net) of the set T is a finite number of points $\mathcal{N} = \{t_1, t_2, \ldots, t_k\} \subset T$ such that every point $t \in T$ is within distortion ϵ of some point of $\mathcal{N}$. In the information theory literature, the points $t_i \in \mathcal{N}$ are called "quantization points", and the process of mapping an arbitrary point $t \in T$ to its closest point in set $\mathcal{N}$ is called *compression* because it allows one to describe each point in set t by just a number from the set $\{1, 2, \ldots, k\}$, *i.e.,* the index of its closest point in $\mathcal{N}$. Let $\hat{X}_i = Y_{t_i}$ for $1 \leqslant i \leqslant k$ be the value of the random process at points in the ϵ-net. The idea of covering is to approximate $X = \sup_{t\in T} Y_t$ by $\max(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_k)$. Since any arbitrary $t \in T$ is close to some point $t_i \in \mathcal{N}$, random variable X too should be close to $\hat{X}_i$ for some i. Once we relate $\sup_{t\in T} Y_t$ to the maximum of finitely many terms $\max(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_k)$, one can use tools such as the union bound to study the latter maximum.

The underlying idea of covering is fairly general. Suppose we have an arbitrary random variable X that is not necessarily arising as the supremum of an underlying random process. We can still apply similar ideas if we "cover" X by a finite collection of random variables $\{\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_k\}$. In this paper, we take a similar approach but with two crucial differences: (i) instead of covering a random variable X, we start off by taking a vector X^m of m i.i.d. repetitions of X, and cover the vector X^m. This technique is known in the information literature as "block-coding" and allows for a certain concentration of measure phenomenon to occur when we let m, the number of repetitions of X to go to infinity. Moreover, it allows to utilize classical results on compression from information theory (ii) instead of covering the entire space as in an ϵ-net, we allow for a vanishing fraction of the space to remain uncovered. This is in line with the information-theoretic notion of covering. More precisely, we cover the subspace in which X^m concentrates on.

To see the idea of block covering in action, let $X_1, \ldots, X_m$ be m i.i.d. repetitions of X. Then, $\mathbb{P}(X \geqslant \Delta)^m = \mathbb{P}(\min_i X_i \geqslant \Delta)$. In order to relate $\min_i X_i$ to an average, we introduce the following distortion function between two sequences.[14]

Definition 13 *Given two sequences* $\boldsymbol{a} = (a_1, \ldots, a_m)$ *and* $\boldsymbol{b} = (b_1, \ldots, b_m)$*, define* $\rho(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \min_i a_i - \frac{1}{m}\sum_i b_i$.

Then, we have the following result (see Appendix E.8 for a proof):

Theorem 14 *Let X be an arbitrary random variable. Take $m \in \mathbb{N}$ and let X^m be its m i.i.d. repetitions. Let $\hat{X}^m(1), \ldots, \hat{X}^m(k)$ be an arbitrary set of $k \in \mathbb{N}$ random variables produced from some arbitrary conditional distribution $Q_{\hat{X}^m(1),\cdots,\hat{X}^m(k)|X^m}$. Then, for any $\epsilon \in \mathbb{R}$,*

14. This distortion function is new and not previously used in the information theory literature to the best of our knowledge.

$$\mathbb{P}(X \geqslant \Delta)^m \leqslant \mathbb{P}\left(\exists j : \frac{1}{m}\sum_{i=1}^{m} \hat{X}_i(j) \geqslant \Delta - \epsilon\right) + \mathbb{P}\Big(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon\Big) \tag{18}$$
$$\leqslant \sum_{j=1}^{k} \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} \hat{X}_i(j) \geqslant \Delta - \epsilon\right) + \mathbb{P}\Big(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon\Big).$$

The random vectors $\hat{X}^m(1), \hat{X}^m(2), \ldots, \hat{X}^m(k)$ represent "quantizations" of the sequences X^m. The term $\mathbb{P}\Big(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon\Big)$ represents probability of excess distortion (with respect to ρ) when covering X^m by $\hat{X}^m(1), \ldots, \hat{X}^m(k)$. The term $\mathbb{P}\Big(\frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon\Big)$ is a tail bound inequality on the quantizations points. We make this more clear by the following example.

Example 1 *Let $k = 1$, $X \sim$ Bernoulli$(1/2)$, Δ=0.5, and ϵ=0. Then, $\mathbb{P}(X \geqslant \Delta) = 0.5$. Let $\hat{X}_i = 0$ with probability one. Then, $\mathbb{P}(\sum \hat{X}_i \geqslant \Delta - \epsilon) = 0$. The term $\rho(X^m, \hat{X}^m)$ is zero if and only if $X_i = 0$ for some i. Thus, $\mathbb{P}(\rho(X^m, \hat{X}^m) > \epsilon) = (1/2)^m$. Hence, we have equality for this example.*

5. Conclusion

In this work, using the source coding literature, we developed a compressibility framework to study the generalization error of the stochastic learning algorithms. This framework allows establishing bounds on the generalization gap in terms of rate-distortion function. Further, our defined compressibility notion makes the connection between several different research approaches in studying the generalization gap, *e.g.* information-theoretic and dimension-based approaches. This study opens up new directions, including: (i) making our bounds computational by applying the numerical methods to compute or bound the rate-distortion function and rate-distortion dimension, (ii) investigating the relation between our bounds and other dimensions-based bounds, by exploiting the relation between rate-distortion dimension and fractal dimensions, *e.g.* correlation dimension, (iii) making the connection between our compressibility framework and PAC-Bayesian approaches (McAllester, 1999),[15] (iv) to establish general bounds on the generalization error by combining rate-distortion theoretic results of this work and the approach of using surrogate hypothesis (Negrea et al., 2020a), and contrariwise, to use the ad-hoc approaches of the latter for our compressibility framework to derive alternative bounds, (v) and finally combining the information-theoretic covering approach, introduced in Section 4.1, with other related techniques such as chaining.

Acknowledgments

This work is partly supported by the French National Research Agency grant ANR-16-CE23-0014 (FBIMATRIX). UŞ's research is supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

15. This relation have been previously established by Blum and Langford (2003) for the compressibility framework of (Littlestone and Warmuth, 1986). The connection of the PAC-Bayesian approaches, particularly when applied for neural networks (MacKay, 1995; Langford and Caruana, 2001; Dziugaite and Roy, 2017; Neyshabur et al., 2018), with our framework also seems promising. In these approaches the propagated error at the output of the network due to small perturbation of the weights are studied. Perturbing W can be seen as letting $\hat{W} = W + N$, where N is an independent noise, and the propagated error as the induced distortion. Then, one needs to properly bound $I(\hat{W}; W)$.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 254–263. PMLR, 10–15 Jul 2018.

Mayank Bakshi and Rakesh K. Bansal. On error exponent in lossy source coding, 2005.

Melih Barsbey, Milad Sefidgaran, Murat A Erdogdu, Gaël Richard, and Umut Şimşekli. Heavy tails in SGD and compressibility of overparametrized neural networks. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Data-dependent coresets for compressing neural networks with applications to generalization bounds. In *International Conference on Learning Representations*, 2019.

Toby Berger. *Rate Distortion Theory and Data Compression*, pages 1–39. Springer Vienna, Vienna, 1975. ISBN 978-3-7091-2928-9.

Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

Avrim Blum and John Langford. Pac-mdl bounds. In *Learning theory and kernel machines*, pages 344–357. Springer, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, May 2020. ISSN 2641-8770.

Yuheng Bu, Weihao Gao, Shaofeng Zou, and Venugopal V. Veeravalli. Population risk improvement with model compression: An information-theoretic approach. *Entropy*, 23(10), 2021.

Alexander Camuto, George Deligiannidis, Murat A. Erdogdu, Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. Fractal structure and generalization properties of stochastic optimization algorithms, 2021.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

Imre Csiszár. Generalized cutoff rates and renyi's information measures. *IEEE Transactions on Information Theory*, 41(1):26–34, 1995.

Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2 edition, 2011.

Paul Warner Cuff, Haim H. Permuter, and Thomas M. Cover. Coordination capacity. *IEEE Transactions on Information Theory*, 56(9):4181–4206, 2010.

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

Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, 2011.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via Rényi-, f-divergences and maximal leakage, 2020.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms, 2020.

Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Daniel M. Roy. Towards a unified information-theoretic framework for generalization. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Te Sun Han. The reliability functions of the general source with fixed-length coding. *IEEE Transactions on Information Theory*, 46(6):2117–2132, 2000.

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms, 2021.

Fredrik Hellstrom and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, Nov 2020. ISSN 2641-8770.

Liam Hodgkinson, Umut Şimşekli, Rajiv Khanna, and Michael W. Mahoney. Generalization properties of stochastic optimizers via trajectory analysis, 2021.

Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. Generalization bounds via distillation. In *International Conference on Learning Representations*, 2021.

Shunsuke Ihara and Masashi Kubo. Error exponent for coding of memoryless gaussian sources with a fidelity criterion. *IEICE Trans. Fundamaentals, A*, 83(10):1891–1897, oct 2000. ISSN 09168508.

Kiminori Iriyama. Probability of error for the fixed-length lossy coding of general sources. *IEEE Transactions on Information Theory*, 51(4):1498–1507, 2005.

Tsutomu Kawabata and Amir Dembo. The rate-distortion dimension of sets and measures. *IEEE Transactions on Information Theory*, 40(5):1564–1572, 1994.

Lorenz Kuhn, Clare Lyle, Aidan N. Gomez, Jonas Rothfuss, and Yarin Gal. Robustness to Pruning Predicts Generalization in Deep Neural Networks. *arXiv:2103.06002 [cs, stat]*, March 2021.

John Langford and Rich Caruana. (not) bounding the true error. *Advances in Neural Information Processing Systems*, 14, 2001.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. *Citeseer*, 1986.

David JC MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469, 1995.

Katalin Marton. Error exponent for source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, 20(2):197–199, 1974.

Mohammad Saeed Masiha, Amin Gohari, Mohammad Hossein Yassaee, and Mohammad Reza Aref. Learning under distribution mismatch and model misspecification. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2912–2917. IEEE, 2021.

David A McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272. PMLR, 2020a.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates, 2020b.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks, 2018.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Erwin Riegler, Helmut Bölcskei, and Günther Koliander. Rate-distortion theory for general sets and measures. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 101–105, 2018.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13 (1):145–147, 1972. ISSN 0097-3165.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Claude E. Shannon. The mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, July 1948.

Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247 – 261, 1972.

Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5138–5151. Curran Associates, Inc., 2020.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 09–12 Jul 2020.

Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura. Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. In *International Joint Conference on Artificial Intelligence*, pages 2839–2846, 2020a.

Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020b.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Yihong Wu and Sergio Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Transactions on Information Theory*, 56(8):3721–3748, 2010.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *NeurIPS*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Appendices

The organization of the appendices is as follows.

- In Appendix A, we introduce the notion of conditional compressibility. Using this concept, we derive several bounds on the generalization performance that recover (and for certain cases improve) some previous bounds by Steinke and Zakynthinou (2020); Harutyunyan et al. (2021); Vapnik (1998).
- In Appendix B, we discuss the Donsker-Varadhan's inequality. The relation with compressibility is shown and a variational representation of the expectation of a random variable is presented.
- In Appendix C, the tail bound on the arbitrary random variable (Theorem 12) is discussed. In particular, a variational representation of the tail probability is presented, which is a key lemma to derive this tail bound.
- In Appendix D, extensions of Theorems 2, 4, and 10 are stated, in addition to some concrete examples of Corollary 11.
- Finally, in Appendix E, proofs of our results are presented.

Type of a sequence Through the appendices, we use the notion of the *type* (Cover and Thomas, 2006). Here, we give its definition. We say that two sequences $x^m, x'^m \in \mathcal{X}^m$ have the same type if their empirical distributions are the same. The type of a sequence x^m is its empirical distribution and is denoted by $\mathcal{T}(x^m)$. An m-type Q_m refers to all sequences of length m whose empirical distributions equal Q_m. Note that for any $x \in \mathcal{X}$, $Q_m(x) = k/m$ where $k \in \{0, 1, \ldots, m\}$. For ease of notation, the type $Q_m(x)$ is simply denoted by $Q(x)$ or $q(x)$, whenever m is known from the context.

Appendix A. Conditional Compressibility

In this section, we introduce conditional compressibility, using concepts from Steinke and Zakynthinou (2020). Building based on this notion, we derive both in expectation and tail bounds that recover and improve over some previous results. Theorem 17 recovers (and potentially improves over) (Steinke and Zakynthinou, 2020, Theorem 1.2.1) and (Harutyunyan et al., 2021, Corollary 2). Corollaries 18 and 22 recover the the in-expectation and tail bound results when a learning algorithm has a bounded VC-dimension (Vapnik, 1998).

Through this section, assume $\mathfrak{Z} \in \mathcal{Z}^{n\times 2}$ be a super-dataset of length $2n$, distributed according to $P_{\mathfrak{Z}} = \mu^{\otimes 2n}$, containing the dataset $S = \mathfrak{Z}_{\mathbf{K}}$ and a ghost dataset $\bar{S} = \mathfrak{Z}_{\bar{\mathbf{K}}}$, where $\mathfrak{Z}_{\mathbf{K}} \coloneqq (\mathfrak{Z}_{1,K_1}, \ldots, \mathfrak{Z}_{n,K_n})$ and $\mathbf{K} \coloneqq (K_1, \ldots, K_n)$ and $\bar{\mathbf{K}} \coloneqq (\bar{K}_1, \ldots, \bar{K}_n)$ are vectors of length n such that each K_i takes values uniformly over $\{1, 2\}$ independent of $\{K_j \colon j \neq i\}$, and $\bar{K}_i \coloneqq \{1, 2\} \backslash K_i$. Denote

$$f(\mathfrak{z}, \mathbf{k}, w) \coloneqq \hat{\mathcal{L}}(\mathfrak{z}_{\bar{\mathbf{k}}}, w) - \hat{\mathcal{L}}(\mathfrak{z}_{\mathbf{k}}, w) = \frac{1}{n}\sum_{j=1}^{n} (-1)^{k_j} (\ell(\mathfrak{z}_{j,1}, w) - \ell(\mathfrak{z}_{j,2}, w)). \tag{19}$$

Let $d_m \coloneqq \mathcal{W}^m \times \hat{\mathcal{W}}^m \times \mathcal{Z}^{2nm} \times \{1,2\}^{nm} \to \mathbb{R}$ be a function, measuring a distortion between m realizations of $f(\mathfrak{z}, \mathbf{k}, w)$ and $f(\mathfrak{z}, \mathbf{k}, \hat{w})$. In particular, we use the following distortion functions:

$$\varphi_m(w^m, \hat{w}^m; \mathfrak{z}^m, \mathbf{k}^m) \coloneqq \min_{j\in[m]} f(\mathfrak{z}_j, \mathbf{k}_j, w_j) - \frac{1}{m}\sum_{i=1}^m f(\mathfrak{z}_i, \mathbf{k}_i, \hat{w}_i), \tag{20}$$

$$\vartheta_m(w^m, \hat{w}^m; \mathfrak{z}^m, \mathbf{k}^m) \coloneqq \frac{1}{m}\sum_{i=1}^m (f(\mathfrak{z}_i, \mathbf{k}_i, w_i) - f(\mathfrak{z}_i, \mathbf{k}_i, \hat{w}_i)), \tag{21}$$

$$\xi_m(w^m, \hat{w}^m; \mathfrak{z}^m, \mathbf{k}^m) \coloneqq \frac{1}{m}\sum_{i=1}^m (|f(\mathfrak{z}_i, \mathbf{k}_i, w_i)| - |f(\mathfrak{z}_i, \mathbf{k}_i, \hat{w}_i)|), \tag{22}$$

where $\mathbf{k}_i = (k_{i,1}, \ldots, k_{i,n})$. In general when $\mathfrak{z}^m = (\mathfrak{z}, \ldots, \mathfrak{z})$, the distortion functions are denoted by $d_m(w^m, \hat{w}^m; \mathfrak{z}, \mathbf{k}^m)$.

Definition 15 *The learning algorithm $\mathcal{A}(S)$ is $(R(\mathfrak{z}), \epsilon(\mathfrak{z}); \{d_m\}_m)$-conditionally compressible for some $R(\mathfrak{z}) \in \mathbb{R}^+$ and $\epsilon(\mathfrak{z}) \in \mathbb{R}$, if for any $\mathfrak{z} \in \mathcal{Z}^{2\times n}$, there exists a sequence of hypothesis books $\{\mathcal{H}_m(\mathfrak{z})\}_{m\in\mathbb{N}}$, $\mathcal{H}_m(\mathfrak{z}) = \{\hat{\mathbf{w}}_j(\mathfrak{z}), j \in [l_m(\mathfrak{z})]\} \subseteq \hat{\mathcal{W}}^m$ such that $l_m(\mathfrak{z}) \leqslant e^{mR(\mathfrak{z})}$ and*

$$\lim_{m\to\infty} \mathbb{P}_{(\mathbf{K}, W|\mathfrak{z})^{\otimes m}}\left(\min_{j\in[l_m(\mathfrak{z})]} d_m(W^m, \hat{\mathbf{w}}_j(\mathfrak{z}); \mathfrak{z}, \mathbf{K}^m) > \epsilon(\mathfrak{z})\right) = 0, \tag{23}$$

where $P_{\mathbf{K},W|\mathfrak{z}} = \frac{1}{2^n} P_{W|\mathfrak{z}\mathbf{K}}$.

The learning algorithm is $(R(\mathfrak{z}), \epsilon(\mathfrak{z}), \delta; \{d_m\}_m)$- exponentially and conditionally compressible for some $\delta > 0$, if in addition to above conditions, the following also holds:

$$\lim_{m\to\infty}\left[-\frac{1}{m}\log\left(\mathbb{P}_{(\mathbf{K}, W|\mathfrak{z})^{\otimes m}}\left(\min_{j\in[l_m(\mathfrak{z})]} d_m(W^m, \hat{\mathbf{w}}_j(\mathfrak{z}); \mathfrak{z}, \mathbf{K}^m) > \epsilon(\mathfrak{z})\right)\right)\right] \geqslant \log(1/\delta). \tag{24}$$

In other words, asymptotically the error probability is bounded by δ^m.

Similar to the unconditional part, we state in expectation and tail bounds.

A.1. Bounds on the expected value of the generalization gap

The first theorem is a bound on the expectation of the generalization performance of the conditionally compressible algorithms.

Theorem 16 *Consider a learning algorithm $\mathcal{A}(S)$ and a bounded loss function $\ell(z, w) \in [0, 1]$.*

i. If $\mathcal{A}(S)$ is $(R(\mathfrak{z}), \epsilon(\mathfrak{z}); \{\vartheta_m\}_m)$-conditionally compressible, then

$$\mathbb{E}[\mathrm{gen}(S, W)] \leqslant \mathbb{E}_{\mathfrak{Z}}\left[\sqrt{\frac{2R(\mathfrak{Z})}{n}} + \epsilon(\mathfrak{Z})\right].$$

ii. If $\mathcal{A}(S)$ is $(R(\mathfrak{z}), \epsilon(\mathfrak{z}); \{|\vartheta_m|\}_m)$-conditionally compressible, then

$$|\mathbb{E}[\mathrm{gen}(S, W)]| \leqslant \mathbb{E}_{\mathfrak{Z}}\left[\sqrt{\frac{2R(\mathfrak{Z})}{n}} + \epsilon(\mathfrak{Z})\right].$$

ii. If $\mathcal{A}(S)$ is $(R(\mathfrak{z}), \epsilon(\mathfrak{z}); \{\xi_m\}_m)$-conditionally compressible, then

$$\mathbb{E}[|\mathrm{gen}(S,W)|] \leqslant \mathbb{E}_{\mathfrak{Z}}\left[\sqrt{\frac{2(R(\mathfrak{Z})+\log(2))}{n}} + \epsilon(\mathfrak{Z})\right].$$

This theorem is proved in Appendix E.9. We use this result to derive a bound on the generalization gap of an arbitrary learning algorithm, in the next theorem. This theorem can be derived from (Steinke and Zakynthinou, 2020, Theorem 1.2) in the same manner as we have proved Theorem 4 using (Xu and Raginsky, 2017, Theorems 1,4). It can be alternatively derived using Theorem 16 and by bounding the conditional compressibility parameters of an arbitrary learning algorithm, similar to the proof of Theorem 3. We omit the proof, as it is similar to the proofs of Theorems 3 and 4.

Theorem 17 *Suppose the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and the loss function $\ell(z,w)$ is bounded in the range $[0,1]$. Consider any auxiliary random variable U[16] defined by the conditional distribution $P_{U|S,W}$ and satisfying $P_{U,S,W} = P_U P_S P_{W|U,S}$[17]. Then, for any $\epsilon \in \mathbb{R}$*

i.

$$\begin{aligned}
|\mathbb{E}_{S,W}[\mathrm{gen}(S,W)]| &\leqslant \mathbb{E}_{\mathfrak{Z},U}\left[\sqrt{\frac{2R_{E,\mathfrak{Z},U}(\epsilon)}{n}} + \epsilon\right],\\
\mathbb{E}_{S,W}[\mathrm{gen}(S,W)] &\leqslant \mathbb{E}_{\mathfrak{Z},U}\left[\sqrt{\frac{2R'_{E,\mathfrak{Z},U}(\epsilon)}{n}} + \epsilon\right],\\
\mathbb{E}_{S,W}[|\mathrm{gen}(S,W)|] &\leqslant \mathbb{E}_{\mathfrak{Z},U}\left[\sqrt{\frac{2(R''_{E,\mathfrak{Z},U}(\epsilon)+\log(2))}{n}} + \epsilon\right],
\end{aligned} \tag{25}$$

where

$$\begin{aligned}
R_{E,\mathfrak{z},u}(\epsilon) &= \inf_{P_{\hat{W}|\mathfrak{z}\mathbf{K},u}} I(\mathbf{K};\hat{W}|\mathfrak{z},u), \quad \textit{such that} \quad \left|\mathbb{E}\Big[f(\mathfrak{z},\mathbf{K},W) - f(\mathfrak{z},\mathbf{K},\hat{W})\Big]\right| \leqslant \epsilon,\\
R'_{E,\mathfrak{z},u}(\epsilon) &= \inf_{P_{\hat{W}|\mathfrak{z}\mathbf{K},u}} I(\mathbf{K};\hat{W}|\mathfrak{z},u), \quad \textit{such that} \quad \mathbb{E}\Big[f(\mathfrak{z},\mathbf{K},W) - f(\mathfrak{z},\mathbf{K},\hat{W})\Big] \leqslant \epsilon,\\
R''_{E,\mathfrak{z},u}(\epsilon) &= \inf_{P_{\hat{W}|\mathfrak{z}\mathbf{K},u}} I(\mathbf{K};\hat{W}|\mathfrak{z},u), \quad \textit{such that} \quad \mathbb{E}\Big[|f(\mathfrak{z},\mathbf{K},W)| - |f(\mathfrak{z},\mathbf{K},\hat{W})|\Big] \leqslant \epsilon.
\end{aligned} \tag{26}$$

The expectations are with respect to $P_{\mathbf{K}}P_{W|\mathfrak{z}\mathbf{K},u}$ and $P_{\mathbf{K}}P_{\hat{W}|\mathfrak{z}\mathbf{K},u}$.

ii.

$$\begin{aligned}
|\mathbb{E}_{S,W}[\mathrm{gen}(S,W)]| &\leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathfrak{Z},U}\Big[\sqrt{2R_{E,\mathfrak{Z},U,i}(\epsilon)} + \epsilon\Big],\\
\mathbb{E}_{S,W}[\mathrm{gen}(S,W)] &\leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathfrak{Z},U}\Big[\sqrt{2R'_{E,\mathfrak{Z},U,i}(\epsilon)} + \epsilon\Big],
\end{aligned}$$

16. Here, U represents the stochasticity of the algorithm. Note that U being a constant is always a valid choice. For further discussions, refer to Appendix D.2.

17. Note that $P_{U,\mathfrak{Z},\mathbf{K},W} = P_{\mathfrak{Z}}P_U P_{W|U,\mathfrak{Z}\mathbf{K}}$.

$$\mathbb{E}_{S,W}[|\text{gen}(S,W)|] \leqslant \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathfrak{Z},U}\Big[\sqrt{2(R''_{E,\mathfrak{Z},U,i}(\epsilon)+\log(2))}+\epsilon\Big], \tag{27}$$

where

$$\begin{aligned}
R_{E,\mathfrak{Z},u,i}(\epsilon) &= \inf_{P_{\hat{W}|\mathfrak{z}_{K_i},u}} I(K_i;\hat{W}|\mathfrak{z}_i,u), \quad \textit{such that} \quad \Big|\mathbb{E}\Big[f(\mathfrak{z}_i,K_i,W)-f(\mathfrak{z}_i,K_i,\hat{W})\Big]\Big| \leqslant \epsilon,\\
R'_{E,\mathfrak{Z},u,i}(\epsilon) &= \inf_{P_{\hat{W}|\mathfrak{z}_{K_i},u}} I(K_i;\hat{W}|\mathfrak{z}_i,u), \quad \textit{such that} \quad \mathbb{E}\Big[f(\mathfrak{z}_i,K_i,W)-f(\mathfrak{z}_i,K_i,\hat{W})\Big] \leqslant \epsilon,\\
R''_{E,\mathfrak{Z},u,i}(\epsilon) &= \inf_{P_{\hat{W}|\mathfrak{z}_{K_i},u}} I(K_i;\hat{W}|\mathfrak{z}_i,u), \quad \textit{such that} \quad \mathbb{E}\Big[|f(\mathfrak{z}_i,K_i,W)|-|f(\mathfrak{z}_i,K_i,\hat{W})|\Big] \leqslant \epsilon,
\end{aligned} \tag{28}$$

where $f(\mathfrak{z}_i,K_i,w) \coloneqq (-1)^{k_i}(\ell(\mathfrak{z}_{i,1},w)-\ell(\mathfrak{z}_{i,2},w))$ *and the expectations are with respect to* $P_{K_i}P_{W|\mathfrak{z}_{i,k_i},u}$ *and* $P_{K_i}P_{\hat{W}|\mathfrak{z}_{i,K_i},u}$.

The above bound trivially recovers (Steinke and Zakynthinou, 2020, Theorem 1.2.1) and (Harutyunyan et al., 2021, Corollary 2) by letting $U=$ Constant, $\epsilon=0$, and $\hat{W}=W$.

Next, we show that we can recover the bound in terms of VC-dimension using the above result.

Corollary 18 *If a learning algorithm has the VC-dimension* d *and the loss function* $\ell(z,w)\in[0,1]$, *then*

$$|\mathbb{E}[\text{gen}(S,W)]| \leqslant \sqrt{\frac{2d\log(2en/d)}{n}}, \quad \mathbb{E}[|\text{gen}(S,W)|] \leqslant \sqrt{\frac{2(d\log(2en/d)+\log(2))}{n}}.$$

The corollary is proved in Appendix E.10.

A.2. Tail bounds on the generalization gap

In this section, we propose tail bounds on the generalization performance using exponentially and conditionally compressibility.

Theorem 19 *If the learning algorithm* $\mathcal{A}(S)$ *is* $(R(\mathfrak{z}),\epsilon(\mathfrak{z}),\delta/2;\{d_{\mathfrak{z},p}\}_m)$*- exponentially and conditionally compressible, then with probability at least* $1-\delta$ *for the bounded loss function* $\ell(z,w)\in[0,1]$,

$$\text{gen}(S,W) \leqslant \sup_{\mathfrak{z}\in\mathcal{Z}^{2n}}\left[\sqrt{\frac{2(R(\mathfrak{z})+\log(2/\delta))}{n}}+\epsilon(\mathfrak{z})\right]+\sqrt{\frac{\log(2/\delta)}{n}}.$$

The theorem is proved in Appendix E.11.

Now, we establish a tail bound on the generalization performance of the arbitrary learning algorithm. For a given distribution $Q_{\mathbf{K},W}$ over $\{1,2\}^n\times\mathcal{W}$, let

$$d_{Q_{\mathbf{K},W}}(\hat{w};\mathfrak{z},\mathbf{k}) \coloneqq \inf_{(\mathbf{k}',w')\in\text{supp}(Q_{\mathbf{K},W})}\big[f(\mathfrak{z},\mathbf{k}',w')\big]-f(\mathfrak{z},\mathbf{k},\hat{w}). \tag{29}$$

Moreover, for a set $\mathcal{U}$[18] and a distribution Q defined over $\{1,2\}^n \times \mathcal{W} \times \mathcal{U}$, define

$$\begin{aligned}\mathfrak{RD}^{\dagger}(\epsilon; Q|_{\mathfrak{z}}) \coloneqq & \inf_{\substack{P_{\hat{W}|\mathfrak{z}_{\mathbf{K}},U}: \\ \mathbb{E}\left[d_{Q_{\mathbf{K},W}}(\hat{W};\mathfrak{z},\mathbf{K})\right]\leqslant\epsilon}} I(\mathbf{K};\hat{W}|U) \\ \leqslant & \inf_{\substack{P_{\hat{W}|\mathfrak{z}_{\mathbf{K}},U}: \\ \mathbb{E}\left[f(\mathfrak{z},\mathbf{K},W)-f(\mathfrak{z},\mathbf{K},\hat{W})\right]\leqslant\epsilon}} I(\mathbf{K};\hat{W}|U) \qquad (30) \\ = & \inf_{\substack{P_{\hat{W}|\mathfrak{z}_{\mathbf{K}},U}: \\ \mathbb{E}\left[\mathrm{gen}(\mathfrak{z}_{\mathbf{K}},W)-\mathrm{gen}(\mathfrak{z}_{\mathbf{K}},\hat{W})\right]\leqslant\epsilon}} I(\mathbf{K};\hat{W}|U), \qquad (31)\end{aligned}$$

where $Q_{\mathbf{K},W}$ is the marginal distribution of $(\mathbf{K}, W)$, the infimum is over all conditional probability distributions (Markov kernels) $P_{\hat{W}|S,U} : \hat{\mathcal{W}} \times \mathcal{S} \times \mathcal{U} \to \mathbb{R}^+$, the expectation and the mutual information are with respect to joint distributions Q and $Q_{U,\mathbf{K}} \times P_{\hat{W}|\mathfrak{z}_{\mathbf{K}},U}$, where $Q_{U,\mathbf{K}}$ is the marginal distribution of $(U, \mathbf{K})$. Then, we have the below tail bound, proved in Appendix E.12.

Theorem 20 *Suppose the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and $\ell(z,w)$ is bounded in the range $[0,1]$. Consider any auxiliary random variable U defined by the conditional distribution $P_{U|S,W}$ and satisfying $P_{U,S,W} = P_U P_S P_{W|U,S}$[19]. Then, for any values of $\{\epsilon(\mathfrak{z})\}_{\mathfrak{z}}$ and $\delta \geqslant 0$, with probability at least $1-\delta$*

$$\mathrm{gen}(S,W) \leqslant \sup_{\mathfrak{z}}\left[\sqrt{\frac{2(R(\mathfrak{z},\delta,\epsilon)+\log(2/\delta))}{n}} + \epsilon(\mathfrak{z})\right] + \sqrt{\frac{\log(2/\delta)}{n}}. \qquad (32)$$

where

$$R(\mathfrak{z},\delta,\epsilon) \coloneqq \sup_{Q:\, D_{KL}(Q\|P_{\mathbf{K},W,U|\mathfrak{z}})\leqslant\log(2/\delta)} \mathfrak{RD}^{\dagger}(\epsilon(\mathfrak{z}); Q|_{\mathfrak{z}}), \qquad (33)$$

where the supremum is over all possible distributions Q over $\{1,2\}^n \times \mathcal{W} \times \mathcal{U}$.

Remark 21 *By considering the exponentially and conditionally compressibility with respect to $\mathbb{P}_{(\mathbf{K},W,\mathfrak{Z})^{\otimes m}}$ rather than $\mathbb{P}_{(\mathbf{K},W|_{\mathfrak{z}})^{\otimes m}}$ in (24), the following result also can be achieved with the assumptions of Theorem 20. For any ϵ and $\delta \geqslant 0$, with probability at least $1-\delta$*

$$\mathrm{gen}(S,W) \leqslant \sqrt{\frac{2(R(\delta,\epsilon)+\log(2/\delta))}{n}} + \epsilon + \sqrt{\frac{\log(2/\delta)}{n}}. \qquad (34)$$

where

$$R(\delta,\epsilon) \coloneqq \sup_{Q:\, D_{KL}(Q\|P_{K,W,\mathfrak{Z},U})\leqslant\log(2/\delta)} \mathbb{E}_{\mathfrak{Z}\sim Q_{\mathfrak{Z}}}\left[\mathfrak{RD}^{\dagger}(\epsilon; Q_{\mathbf{K},W,U|\mathfrak{Z}}|\mathfrak{Z})\right], \qquad (35)$$

where the supremum is over all possible distributions Q over $\{1,2\}^n \times \mathcal{W} \times \mathcal{Z}^{2n} \times \mathcal{U}$, $Q_{\mathbf{K},W,U|\mathfrak{Z}}$ is the conditional distribution of $(\mathbf{K}, W, U)$ given $\mathfrak{Z}$, and $Q_{\mathfrak{Z}}$ is the marginal distribution of $\mathfrak{Z}$.

18. As mentioned before, U represents the stochasticity of the algorithm. For further discussions, refer to Appendix D.2.
19. Note that $P_{\mathfrak{Z},\mathbf{K},W,U} = P_{\mathfrak{Z}} P_{\mathbf{K}} P_U P_{W|U,\mathfrak{Z}_{\mathbf{K}}} = \frac{1}{2^n}\mu^{\otimes 2n} P_U P_{W|U,\mathfrak{Z}_{\mathbf{K}}}$.

Finally, we use Theorem 20 to recover the generalization bound for the algorithms having a bounded VC-dimension Vapnik (1998).

Corollary 22 *If a learning algorithm has VC-dimension d and the loss function $\ell(z,w) \in [0,1]$, then with probability at least $1-\delta$*

$$\mathrm{gen}(S,W) \leqslant \sqrt{\frac{2(d\log(2en/d)+\log(2/\delta))}{n}} + \sqrt{\frac{\log(2/\delta)}{n}}.$$

The corollary is proved in Appendix E.13.

Appendix B. On the Donsker-Varadhan's Inequality

The Donsker-Varadhan's identity implies that for arbitrary distributions $p(x)$ and $q(x)$ on a set $\mathcal{X}$ and for any arbitrary function $\Phi : \mathcal{X} \to \mathbb{R}$ we have

$$D_{KL}(q\|p) \geqslant \mathbb{E}_{X\sim q}[\Phi(X)] - \log\Big(\mathbb{E}_{X\sim p}\Big[e^{\Phi(X)}\Big]\Big). \tag{36}$$

In this appendix, we first show that this inequality can be proved using a compressibility approach for a finite set $\mathcal{X}$. Then, we also derive a lemma based on (36) that is used to derive a tail bound on an arbitrary random variable.

B.1. Donsker-Varadhan's inequality via compression

Take some arbitrary function $\Phi : \mathcal{X} \to \mathbb{R}$. Generate 2^{mR} sequences

$$X^m(1), X^m(2), \ldots, X^m(2^{mR}),$$

where $X^m(i) = (X_1(i), \ldots, X_m(i))$, in an i.i.d. fashion from $p(x)$, *i.e.* each $X_j(i) \sim p$ for $i \in [2^{mR}], j \in [m]$, independent of other instances. Consider the expression

$$\mathbb{E}\left[\max_k \sum_{i=1}^m \Phi(X_i(k))\right].$$

On the one hand,

$$e^{\mathbb{E}\left[\max_k \sum_i \Phi(X_i(k))\right]} \leqslant \mathbb{E}\Big[e^{\max_k \sum_i \Phi(X_i(k))}\Big] \leqslant \mathbb{E}\left[\sum_k e^{\sum_i \Phi(X_i(k))}\right] = 2^{mR}\Big(\mathbb{E}_{X\sim p}\Big[e^{\Phi(X)}\Big]\Big)^m.$$

Therefore,

$$\mathbb{E}\left[\max_k \sum_i \Phi(X_i(k))\right] \leqslant mR + m\log\Big(\mathbb{E}_{X\sim p}\Big[e^{\Phi(X)}\Big]\Big). \tag{37}$$

On the other hand, let γ_m be the probability that at least one of the sequences $X^m(k)$ for some k will have type $q(x)$. It is known that (for example by using Cover and Thomas (2006, Theorem 11.1.4))

$\gamma_m \to 1$ as m tends to infinity if $R > D_{KL}(q\|p)$. Under the event that the sequence $X^m(k)$ has type $q(x)$, $\sum_i \Phi(X_i(k))$ equals $m\mathbb{E}_{X\sim q}[\Phi(X)]$. Thus,

$$\mathbb{E}\left[\max_k \sum_i \Phi(X_i(k))\right] \geqslant m\gamma_m \mathbb{E}_{X\sim q}[\Phi(X)] + m(1-\gamma_m)\min_x \Phi(x). \tag{38}$$

From (37) and (38) and by letting m tend to infinity, we obtain

$$D_{KL}(q\|p) + \log\Big(\mathbb{E}_{X\sim p}\Big[e^{\Phi(X)}\Big]\Big) \geqslant \mathbb{E}_{X\sim q}[\Phi(X)].$$

This yields the desired inequality.

B.2. Variational representation of $\mathbb{E}[X]$

In this subsection, we state a variational lemma on $\mathbb{E}[X]$, used in proof of Theorem 12. The lemma is proved in Appendix E.14, by using (36).

Lemma 23 *For every distribution ν, we have*

$$\mathbb{E}_\nu[X] = \frac{1}{\lambda}\inf_\mu \Big[D_{KL}(\nu\|\mu) + \log \mathbb{E}_\mu\Big[e^{\lambda X}\Big]\Big]. \tag{39}$$

Appendix C. Tail Bound on an Arbitrary Random Variable

The key to the proof of the tail bound in Theorem 12 is a variational representation of the tail probability, stated in the next lemma.

Lemma 24 *Let ϵ be an arbitrary real number. For any arbitrary distribution ν_X on $\mathbb{R}$, let $\mathcal{P}(\nu_X)$ denote the set of distributions $p_{\hat{X}}$ on $\mathbb{R}$ for which*

$$\left[\inf_{x\in \mathrm{supp}(\nu_X)} x\right] - \mathbb{E}\Big[\hat{X}\Big] \leqslant \epsilon.$$

Let $X \sim \mu_X$ where μ_X is an arbitrary distribution on the sample space $\mathcal{X} \subseteq \mathbb{R}$. Then, for any $\Delta \in \mathbb{R}$ we have[20]

$$\log \mathbb{P}_{X\sim\mu_X}(X \geqslant \Delta) = \sup_{\nu_X \ll \mu_X} \inf_{p_{\hat{X}}\in\mathcal{P}(\nu_X),\ \lambda\geqslant 0} \left\{-D_{KL}(\nu_X\|\mu_X) - \lambda\left[\Delta - \epsilon - \mathbb{E}_p[\hat{X}]\right]_+\right\}.$$

We give two proofs for this lemma. The first proof is provided in Appendix E.15.1. We give also an alternative proof (in the inequality form) when $\mathcal{X}$ and $\mathcal{S}$ are finite sets. This proof illustrates the connections between the tail bound and compression. To this end, consider the distortion function ρ defined in Definition 13. Note that φ_m can be expressed in terms of this distortion function, *i.e.* $\varphi_m(w^m, \hat{w}^m; s^m) = \rho\Big(\{\mathrm{gen}(s_i, w_i)\}_{i\in[m]}, \{\mathrm{gen}(s_i, \hat{w}_i)\}_{i\in[m]}\Big)$. To establish the tail bound, first we upper bound it in terms of the tail of some quantizations of X^m and the probability of covering X^m by this quanitzation points. This is exactly the bound established in Theorem 14. Indeed, Theorem 14 shows the connection between the tail bound and compression. Note that Theorem 14 also holds if the conditions $\geqslant \Delta - \epsilon$ and $> \epsilon$ are replaced by conditions $> \Delta - \epsilon$ and $\geqslant \epsilon$ respectively in (18). The rest of proof applies some information-theoretic techniques, as detailed in Appendix E.15.2.

20. For $a \in \mathbb{R}$, $[a]_+ \coloneqq \max(0, a)$.

Appendix D. Other Results

In this section, we state some further obtained results.

D.1. Extension of the in expectation bound

In a similar manner as (8), let $\xi_m(w^m, \hat{w}^m; s^m) \coloneqq \frac{1}{m}\sum_{i=1}^m(|\operatorname{gen}(s_i, w_i)| - |\operatorname{gen}(s_i, \hat{w}_i)|)$. Here, we state the extended version of Theorem 2.

Theorem 2 *Consider a learning algorithm $\mathcal{A}(S)$ and suppose that $\mathbb{E}_{S,W}[|\operatorname{gen}(S,W)|] < \infty$ and for all $w \in \mathcal{W}$, $\ell(Z,w)$ is σ-subgaussian.*

i. If $\mathcal{A}(S)$ is $(R, \epsilon, \{\vartheta_m\}_m)$-compressible, then $\mathbb{E}[\operatorname{gen}(S,W)] \leqslant \sqrt{2\sigma^2 R/n} + \epsilon$.

ii. If $\mathcal{A}(S)$ is $(R, \epsilon, \{|\vartheta_m|\}_m)$-compressible, then $|\mathbb{E}[\operatorname{gen}(S,W)]| \leqslant \sqrt{2\sigma^2 R/n} + \epsilon$.

iii. If $\mathcal{A}(S)$ is $(R, \epsilon, \{\xi_m\}_m)$-compressible, then $\mathbb{E}[\,|\operatorname{gen}(S,W)|\,] \leqslant \sqrt{2\sigma^2(R+\log(2))/n} + \epsilon$.

All above expectations are with respect to $P_{S,W}$.

Next, we state the extended version of Theorem 4.

Theorem 4 *Assume that the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and for all $w \in \mathcal{W}$, $\ell(Z,w)$ is σ-subgaussian. Then, for any $\epsilon \in \mathbb{R}$,*

$$
\begin{aligned}
|\mathbb{E}[\operatorname{gen}(S,W)]| &\leqslant \sqrt{\frac{2\sigma^2 R_E(\epsilon)}{n}} + \epsilon, \\
\mathbb{E}[\operatorname{gen}(S,W)] &\leqslant \sqrt{\frac{2\sigma^2 R'_E(\epsilon)}{n}} + \epsilon, \\
\mathbb{E}[\,|\operatorname{gen}(S,W)|\,] &\leqslant \sqrt{\frac{2\sigma^2 (R''_E(\epsilon) + \log(2))}{n}} + \epsilon,
\end{aligned} \tag{40}
$$

where

$$
\begin{aligned}
R_E(\epsilon) &= \inf_{P_{\hat{W}|S}} I(S;\hat{W}), \quad \textit{such that} \quad \left|\mathbb{E}\Big[\operatorname{gen}(S,W) - \operatorname{gen}(S,\hat{W})\Big]\right| \leqslant \epsilon, \\
R'_E(\epsilon) &= \inf_{P_{\hat{W}|S}} I(S;\hat{W}), \quad \textit{such that} \quad \mathbb{E}\Big[\operatorname{gen}(S,W) - \operatorname{gen}(S,\hat{W})\Big] \leqslant \epsilon, \\
R''_E(\epsilon) &= \inf_{P_{\hat{W}|S}} I(S;\hat{W}), \quad \textit{such that} \quad \mathbb{E}\Big[|\operatorname{gen}(S,W)| - |\operatorname{gen}(S,\hat{W})|\Big] \leqslant \epsilon.
\end{aligned} \tag{41}
$$

All expectations in above are with respect to $P_{S,W}$ and $P_S \times P_{\hat{W}|S}$.

This theorem can be trivially extended to the case where we have access to an internal randomness U of the algorithm, as defined in Appendix D.2. For example, by using $\mathbb{E}[\operatorname{gen}(S,W)] = \mathbb{E}_U\mathbb{E}_{S,W|U}[\operatorname{gen}(S,W)]$, it can be shown that $\mathbb{E}[\operatorname{gen}(S,W)] \leqslant \mathbb{E}_U[\sqrt{2\sigma^2 R_{E,U}/n}] + \epsilon$, where $R_{E,u}(\epsilon) \coloneqq \inf I(S;\hat{W}|U=u)$, in which the infimum is over all Markov kernels $P_{\hat{W}|S,u}$ such that $\mathbb{E}\Big[\operatorname{gen}(S,W) - \operatorname{gen}(S,\hat{W})\Big] \leqslant \epsilon$.

In the following, Theorem 4 is extended similarly to Bu et al. (2020). The proof trivially follows from the relation $\mathbb{E}[\operatorname{gen}(S,W)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\operatorname{gen}(\{Z_i\},W)]$ and Theorem 4, where $\mathbb{E}[\operatorname{gen}(\{z_i\},w)] = \mathcal{L}(w) - \ell(z_i,w)$.

Theorem 25 *Suppose the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and for all $w \in \mathcal{W}$, $\ell(Z,w)$ is σ-subgaussian. Then, for any $\epsilon \in \mathbb{R}$,*

$$|\mathbb{E}[\operatorname{gen}(S,W)]| \leqslant \frac{1}{n}\sum_{i=1}^{n}\left[\sqrt{2\sigma^2 R_{E,i}(\epsilon)}\right] + \epsilon, \tag{42}$$

$$\mathbb{E}[\operatorname{gen}(S,W)] \leqslant \frac{1}{n}\sum_{i=1}^{n}\left[\sqrt{2\sigma^2 R'_{E,i}(\epsilon)}\right] + \epsilon, \tag{43}$$

$$\mathbb{E}[\,|\operatorname{gen}(S,W)|\,] \leqslant \frac{1}{n}\sum_{i=1}^{n}\left[\sqrt{2\sigma^2 (R''_{E,i}(\epsilon)+\log(2))}\right] + \epsilon, \tag{44}$$

where the expectation is with respect to $P_{S,W}$ and

$$R_{E,i}(\epsilon) = \inf_{P_{\hat{W}|Z_i}} I(Z_i;\hat{W}), \quad \textit{such that} \quad \left|\mathbb{E}\Big[\operatorname{gen}(\{Z_i\},W) - \operatorname{gen}(\{Z_i\},\hat{W})\Big]\right| \leqslant \epsilon, \tag{45}$$

$$R'_{E,i}(\epsilon) = \inf_{P_{\hat{W}|Z_i}} I(Z_i;\hat{W}), \quad \textit{such that} \quad \mathbb{E}\Big[\operatorname{gen}(\{Z_i\},W) - \operatorname{gen}(\{Z_i\},\hat{W})\Big] \leqslant \epsilon, \tag{46}$$

$$R''_{E,i}(\epsilon) = \inf_{P_{\hat{W}|Z_i}} I(Z_i;\hat{W}), \quad \textit{such that} \quad \mathbb{E}\Big[|\operatorname{gen}(\{Z_i\},W)| - |\operatorname{gen}(\{Z_i\},\hat{W})|\Big] \leqslant \epsilon, \tag{47}$$

where $\mathbb{E}[\operatorname{gen}(\{z_i\},w)] = \mathcal{L}(w) - \ell(z_i,w)$ and the expectations are with respect to $P_{Z_i,W}$ and $P_{Z_i} \times P_{\hat{W}|Z_i}$.

Letting $\epsilon = 0$ and $\hat{W} = W$, this theorem recovers (and potentially improves over) (Bu et al., 2020, Proposition 1).

D.2. Extension of the tail bound

It has been already shown by Harutyunyan et al. (2021) that taking into account the stochasticity of the algorithm could yield tighter bounds on the expectation of the generalization gap. Here, we apply a similar idea for the tail bound. To this end, we represent partial or full stochasticity of the algorithm which is independent of the dataset by $U \in \mathcal{U}$. This means that the hypothesis is chosen according to $P_{W|S,U}$ (deterministically or randomly). Having this stochasticity available, we can make our compression more efficient, by letting the hypothesis books in Definition 1 depend on U as well, *i.e.* for each arbitrary distribution Q defined over $\mathcal{U}$, we choose a sequence of hypothesis books $\mathcal{H}_m(Q) = \{\hat{\mathbf{w}}_j(Q), j \in [l_m(Q)]\} \subseteq \hat{\mathcal{W}}^m$, such that $l_m(Q) \leqslant e^{mR(Q)}$ and

$$\lim_{m\to\infty} -\frac{1}{m}\log\left(\mathbb{P}_{(U,S,W)^{\otimes m}}\left(\min_{j\in[l_m(\hat{P}_{U^m})]} \varphi_m\Big(W^m, \hat{\mathbf{w}}_j(\hat{P}_{U^m}); S^m\Big) > \epsilon\right)\right) \geqslant \log(1/\delta), \tag{48}$$

where $\hat{P}_{U^m}$ is the empirical distribution of U^m. Then, it can be shown that R in Theorem 9 can be replaced by $\sup_Q R(Q)$, where the supremum is over all Q such that $D_{KL}(Q\|P_U) \leqslant \log(1/\delta)$.

In order to define the extended version of Theorem 10, we need to define an extended definition of $\mathfrak{RD}^*(\epsilon;Q)$, that takes U also into account. For a distribution Q defined over $\mathcal{S}\times\mathcal{W}\times\mathcal{U}$, let

$$\mathfrak{RD}^*(\epsilon;Q) \coloneqq \inf_{\substack{P_{\hat{W}|S,U}:\\ \mathbb{E}\left[d_{Q_{S,W}}(\hat{W};S)\right]\leqslant\epsilon}} I(S;\hat{W}|U) \leqslant \inf_{\substack{P_{\hat{W}|S,U}:\\ \mathbb{E}\left[\operatorname{gen}(S,W)-\operatorname{gen}(S,\hat{W})\right]\leqslant\epsilon}} I(S;\hat{W}|U), \tag{49}$$

where $Q_{S,W}$ is the marginal distribution of (S, W), the infimum is over all Markov kernels $P_{\hat{W}|S,U} : \hat{\mathcal{W}} \times \mathcal{S} \times \mathcal{U} \to \mathbb{R}^+$, the expectation and the mutual information are with respect to joint distributions Q and $Q_{S,U} \times P_{\hat{W}|S,U}$, where $Q_{S,U}$ is the marginal distribution of (S, U). Note that letting U being a constant, (49) will be reduced to (14). Now, we state an extended version of Theorem 10, proved in Appendix E.6.

Theorem 10 *Suppose that the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and for all $w \in \mathcal{W}$, $\ell(Z, w)$ is σ-subgaussian. Consider any auxiliary random variable U defined by $P_{U|S,W}$ and satisfying $P_{U,S,W} = P_U P_S P_{W|U,S}$. Then, for every $\epsilon \in \mathbb{R}$ and $\delta \geqslant 0$, with probability at least $1 - \delta$,*

$$\operatorname{gen}(S, W) \leqslant \sqrt{\frac{2\sigma^2(R_p(\delta, \epsilon) + \log(1/\delta))}{n}} + \epsilon, \; R_p(\delta, \epsilon) \coloneqq \sup_{Q\colon\, D_{KL}(Q\|P_{S,W,U}) \leqslant \log(1/\delta)} \mathfrak{RD}^*(\epsilon; Q),$$

where the supremum is over all possible distributions Q over $\mathcal{S} \times \mathcal{W} \times \mathcal{U}$.

Note that the above bound holds for any U that satisfies the assumptions of the theorem and U being a constant is always valid choice. By the choice of $U =$ Constant, this extended version becomes the same as the original one, stated in Section 3.3.

As a special case, when S and W are independent, the above theorem, by choosing $U = W$, results that with probability $1 - \delta$, $\operatorname{gen}(S, W) \leqslant \sqrt{2\sigma^2 \log(1/\delta)/n}$. This bound is equal to the one obtainable by direct application of Hoeffding's inequality. However, we may not be able to achieve this bound using Theorem 10 with constant U. Since, while for example $I(S; W) = 0$ under $P_{S,W}$, it may not be equal to zero under distribution Q, where $D_{KL}(Q\|P_{S,W}) \leqslant \log(1/\delta)$; as under distribution Q, random variables S and W might be (weakly) dependent.

D.3. Examples for Lipschitz loss

In the following, we show some consequences of Corollary 11.

Corollary 26 *Suppose that the loss function $\ell(Z, w)$ is σ-subgaussian for any $w \in \mathcal{W}$.*

i. [ε-net covering] Let $\mathcal{W} = \mathbb{R}^d$ and let W with probability one take value in the d-dimensional ball $\mathcal{V}_d = \{w \in \mathbb{R}^d \colon \|w\| \leqslant r_0\}$ and suppose that for every $z, w, \hat{w}$, $|\ell(z, w) - \ell(z, \hat{w})| \leqslant \mathfrak{L}\|w - \hat{w}\|$. Then, for every $\delta > 0$, with probability at least $1 - \delta$,

$$\operatorname{gen}(S, W) \leqslant \min_{\epsilon \geqslant 0} \left[\sqrt{\frac{2\sigma^2(d \log(2r_0/\epsilon) + \log(1/\delta))}{n}} + 2\mathfrak{L}\epsilon \right].$$

In particular, for $n \geqslant 16$, with probability at least $1 - e^{-d/2}$, we have $\operatorname{gen}(S, W) \leqslant (4r_0\mathfrak{L} + \sigma\sqrt{d})\sqrt{\log(n)/n}$.

ii. Suppose that $W \in \{0, 1\}^d$ is composed of d i.i.d. elements distributed according to Bernoulli distribution with an unknown parameter $\mathbb{P}(W_i = 1) = p$ and $|\ell(z, w) - \ell(z, \hat{w})| \leqslant \mathfrak{L} d_H(w, \hat{w})$, where d_H is the Hamming distance.[21] Then, for every $\delta > 0$, with probability at least $1 - \delta$,

$$\operatorname{gen}(S, W) \leqslant \min_{0 \leqslant \epsilon \leqslant d} \left[\sqrt{\frac{2\sigma^2[d \log(2) - d h_b(\epsilon/d) + \log(1/\delta)]}{n}} + 2\mathfrak{L}\epsilon \right],$$

21. For binary vectors $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$, $d_H(x, y) \coloneqq \sum_{i=1}^d \mathbb{1}_{\{x_i \neq y_i\}}$.

where $h_b(\cdot)$ is the binary entropy function, i.e. $h_b(p) = -p\log(p) - (1-p)\log(1-p)$*, for* $p \in [0,1]$ *and* $0\log(0) = 0$ *by convention.*

iii. Suppose that $W \in \mathbb{R}^d$ is composed of d i.i.d. elements distributed according to the two-sided exponential distribution $p(w_i) = \frac{\lambda}{2}e^{-\lambda|w_i|}$, $i \in [d]$ and $|\ell(z,w) - \ell(z,\hat{w})| \leqslant \mathfrak{L}\|w - \hat{w}\|_1$. Then, for every $\delta > 0$, with probability at least $1-\delta$,

$$\operatorname{gen}(S,W) \leqslant \min_{\epsilon \geqslant 0}\left[\sqrt{\frac{2\sigma^2(dR'(\epsilon,\delta) + \log(1/\delta))}{n}} + 2\mathfrak{L}\epsilon\right],$$

where $R'(\epsilon,\delta)$ is determined by

$$\log(1/\delta) = \alpha\lambda - 1 - \log(\alpha\lambda),$$

in which $\alpha \coloneqq \epsilon\exp(R'(\epsilon,\delta))/d$.

iv. Suppose that $W \in \mathbb{R}^d$ is composed of d i.i.d. elements distributed according to the normal distribution $\mathcal{N}(0,\sigma_N^2)$ and $|\ell(z,w) - \ell(z,\hat{w})| \leqslant \mathfrak{L}\|w - \hat{w}\|_2^2$. Then, for every $\delta > 0$, with probability at least $1-\delta$,

$$\operatorname{gen}(S,W) \leqslant \min_{\epsilon \geqslant 0}\left[\sqrt{\frac{\sigma^2(d\log(\max(d\alpha^2/\epsilon, 1)) + 2\log(1/\delta))}{n}} + 2\mathfrak{L}\epsilon\right],$$

where $\alpha \geqslant \sigma$ is determined by

$$\log(1/\delta) = \frac{1}{2}\left(\frac{\alpha^2}{\sigma_N^2} - 1 - \log\left(\frac{\alpha^2}{\sigma_N^2}\right)\right).$$

The corollary is proved in Appendix E.16.

Appendix E. Proofs

In this section, we present the proofs of all our results, in the order of their appearances in the paper.

E.1. Proof of Theorem 2

Here we state the proof for the long version of the theorem, stated in Appendix D.1. Note that as defined in that appendix, $\xi_m(w^m, \hat{w}^m; s^m) \coloneqq \frac{1}{m}\sum_{i=1}^m(|\operatorname{gen}(s_i, w_i)| - |\operatorname{gen}(s_i, \hat{w}_i)|)$.

Before stating the proof, we show that having condition (8) for the distortion functions $\{\vartheta_m, |\vartheta_m|, \xi_m\}$ guarantees that the expectation of the difference of the original and compressed algorithms does not exceed ϵ.

Lemma 27 *If $\mathbb{E}[|\operatorname{gen}(S,W)|] < \infty$, then for $d_m \in \{\vartheta_m, |\vartheta_m|, \xi_m\}$ condition (7) yields*

$$\lim_{m\to\infty} \mathbb{E}_{(S,W)^{\otimes m}}\left[\min_{j\in[l_m]} d_m(W^m, \hat{\mathbf{w}}_j; S^m)\right] \leqslant \epsilon. \tag{50}$$

The above lemma is proved in Appendix E.17. Now, we proceed with the proof of the extended version of Theorem 2, appeared in Appendix D.1.

Proof

Part i. Suppose that for each (s^m, w^m), $\hat{\mathbf{w}}(s^m, w^m) \coloneqq \hat{\mathbf{w}}_j$ where $j = \underset{j\in[l_m]}{\arg\min}\, \vartheta_m(w^m, \hat{\mathbf{w}}_j; s^m)$, which will be denoted by $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_m)$ for simplicity. Then,

$$
\begin{aligned}
&\mathbb{E}_{(S,W)}[\operatorname{gen}(S,W)] \\
&\quad = \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\sum_{i=1}^{m} \operatorname{gen}(S_i, W_i)\right] \\
&\quad = \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\sum_{i=1}^{m} \operatorname{gen}(S_i, W_i) - \operatorname{gen}(S_i, \hat{W}_i)\right] + \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\sum_{i=1}^{m} \operatorname{gen}(S_i, \hat{W}_i)\right] \\
&\quad \overset{(a)}{\leqslant} \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\sum_{i=1}^{m} \operatorname{gen}(S_i, \hat{W}_i)\right] + \epsilon + \varepsilon_m \\
&\quad \leqslant \frac{1}{m}\mathbb{E}_{S^{\otimes m}}\left[\max_{j\in[l_m]} \sum_{i=1}^{m} \operatorname{gen}(S_i, \hat{w}_{j,i})\right] + \epsilon + \varepsilon_m \\
&\quad \overset{(b)}{\leqslant} \frac{1}{m}\sqrt{\frac{2\sigma^2 m \log(l_m)}{n}} + \epsilon + \varepsilon_m \\
&\quad \overset{(c)}{\leqslant} \sqrt{\frac{2\sigma^2 R}{n}} + \epsilon + \varepsilon_m,
\end{aligned}
$$

where (a) is by Lemma 27, (b) is derived since $\sum_{i=1}^{m} \operatorname{gen}(S_i, \hat{w}_{j,i})$ is $\sigma\sqrt{m/n}$-subgaussian, and (c) is obtained by bounding $l_m \leqslant e^{mR}$. Taking the limit for $m \to \infty$ completes the proof.

Part ii. Similarly, let $\hat{\mathbf{w}}(s^m, w^m) \coloneqq \hat{\mathbf{w}}_j$ where $j = \underset{j\in[l_m]}{\arg\min}\, |\vartheta_m(w^m, \hat{\mathbf{w}}_j; s^m)|$. Then, we have

$$
\begin{aligned}
&\left|\mathbb{E}_{(S,W)}[\operatorname{gen}(S,W)]\right| \\
&\quad \leqslant \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\left|\sum_{i=1}^{m} \operatorname{gen}(S_i, W_i)\right|\right] \\
&\quad \leqslant \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\left|\sum_{i=1}^{m} \operatorname{gen}(S_i, W_i) - \operatorname{gen}(S_i, \hat{W}_i)\right|\right] + \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\left|\sum_{i=1}^{m} \operatorname{gen}(S_i, \hat{W}_i)\right|\right] \\
&\quad \leqslant \frac{1}{m}\mathbb{E}_{S^{\otimes m}}\left[\max_{j\in[l_m]} \left|\sum_{i=1}^{m} \operatorname{gen}(S_i, \hat{w}_{j,i})\right|\right] + \epsilon + \varepsilon_m \\
&\quad \leqslant \frac{1}{m}\sqrt{\frac{2\sigma^2 m \log(2l_m)}{n}} + \epsilon + \varepsilon_m \\
&\quad \leqslant \sqrt{\frac{2\sigma^2 (R + \log(2)/m)}{n}} + \epsilon + \varepsilon_m.
\end{aligned}
$$

Taking the limit for $m \to \infty$ completes the proof.

Part iii. Similarly, let $\hat{\mathbf{w}}(s^m, w^m) \coloneqq \hat{\mathbf{w}}_j$ where $j = \underset{j\in[l_m]}{\arg\min}\, \xi_m(w^m, \hat{\mathbf{w}}_j; s^m)$. Then, we have

$$
\begin{aligned}
&\mathbb{E}_{(S,W)}\Big[\big|\operatorname{gen}(S,W)\big|\Big] \\
&\quad= \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\sum_{i=1}^{m}\big|\operatorname{gen}(S_i,W_i)\big| - \big|\operatorname{gen}(S_i,\hat{W}_i)\big|\right] + \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left[\sum_{i=1}^{m}\big|\operatorname{gen}(S_i,\hat{W}_i)\big|\right] \\
&\quad\leqslant \frac{1}{m}\mathbb{E}_{S^{\otimes m}}\left[\max_{j\in[l_m]}\sum_{i=1}^{m}\big|\operatorname{gen}(S_i,\hat{w}_{j,i})\big|\right] + \epsilon + \varepsilon_m \\
&\quad\leqslant \frac{1}{m}\sqrt{\frac{2\sigma^2 m\log(2^m l_m)}{n}} + \epsilon + \varepsilon_m \\
&\quad\leqslant \sqrt{\frac{2\sigma^2(R+\log(2))}{n}} + \epsilon + \varepsilon_m.
\end{aligned}
$$

Taking the limit for $m \to \infty$ completes the proof. ■

E.2. Proof of Theorem 3

Proof Fix ϵ and $\nu_1, \nu_2 > 0$. Assume that there exists a $\hat{W} \in \hat{\mathcal{W}}$ defined by the conditional distribution $P_{\hat{W}|S}$, such that $\Big|\mathbb{E}\Big[\operatorname{gen}(S,W) - \operatorname{gen}(S,\hat{W})\Big]\Big| \leqslant \epsilon$. It is sufficient to show $R_E(\epsilon) \leqslant I(S;\hat{W}) + \nu_1$.

Denote the empirical distribution of a sequence x^m by $\hat{P}_{x^m}$, *i.e.*

$$
\hat{P}_{x^m}(x) \coloneqq \frac{\#i \in [m] \colon x_i = x}{m}. \tag{51}
$$

Then, using the proof of (Cuff et al., 2010, Theorem 3), there exists a required sequence of $\{\mathcal{H}_m\}_{m\in\mathbb{N}}$ such that $|\mathcal{H}_m| \leqslant e^{m(I(S;\hat{W})+\nu_1)}$ and such that for each S^m, a vector $\hat{W}^m(S^m) \in \mathcal{H}_m$, that we denote for ease of notations as $\hat{W}^m$, can be chosen such that

$$
\|\hat{P}_{(S^m,\hat{W}^m)}(s,\hat{w}) - P_{(S,\hat{W})}(s,\hat{w})\|_{TV} \xrightarrow{p} 0. \tag{52}
$$

where $\xrightarrow{p}$ means convergence in probability and *TV* denotes the total variation distance between two distributions (Cuff et al., 2010, Definition 4). Moreover, by strong law of large numbers, for m independent instances (S_i, W_i) chosen according to $P_{S,W}$, we have

$$
\|\hat{P}_{(S^m,W^m)}(s,w) - P_{S,W}(s,w)\|_{TV} \xrightarrow{p} 0. \tag{53}
$$

This yields

$$
\begin{aligned}
\left|\frac{1}{m}\sum_{i=1}^{m}\Big[\operatorname{gen}(S_i,W_i) - \operatorname{gen}(S_i,\hat{W}_i)\Big]\right| &\leqslant \Big|\mathbb{E}\Big[\operatorname{gen}(S,W) - \operatorname{gen}(S,\hat{W})\Big]\Big| + \epsilon_m &(54)\\
&\leqslant \epsilon + \varepsilon_m, &(55)
\end{aligned}
$$

where ε_m vanishes as $m \to \infty$. Now,

$$
\begin{aligned}
\mathbb{P}_{(S,W)^{\otimes m}}\Big(\vartheta_m(W^m,\hat{W}^m;S^m) \geqslant \epsilon + \nu_2\Big) &\leqslant \mathbb{P}_{(S,W)^{\otimes m}}(\epsilon + \varepsilon_m \geqslant \epsilon + \nu_2) \\
&\to 0,
\end{aligned}
$$

where the last line is when $m \to \infty$. This completes the proof. ■

E.3. Proof of Theorem 4

Proof We show

$$|\mathbb{E}[\mathrm{gen}(S,W)]| \leqslant \sqrt{\frac{2\sigma^2 R_E(\epsilon)}{n}} + \epsilon, \tag{56}$$

and the proof for the rest of bounds in (40) is similar. Consider any Markov kernel $P_{\hat{W}|S}$ that satisfies $\left|\mathbb{E}\Big[\mathrm{gen}(S,W) - \mathrm{gen}(S,\hat{W})\Big]\right| \leqslant \epsilon$. Then,

$$\begin{aligned}|\mathbb{E}[\mathrm{gen}(S,W)]| \leqslant& \left|\mathbb{E}\Big[\mathrm{gen}(S,\hat{W})\Big]\right| + \epsilon \\ \leqslant& \sqrt{\frac{2\sigma^2 I(S;\hat{W})}{n}} + \epsilon,\end{aligned}$$

where the last step is deduced from (Xu and Raginsky, 2017, Theorem 1). This completes the proof. ■

E.4. Proof of Corollary 7

Proof Let n_0' be large enough such that for $n \geqslant n_0'$ and $\epsilon \coloneqq 2\mathfrak{L}/\sqrt{n\mathfrak{L}^2} = 2/\sqrt{n}$,

$$\mathfrak{RD}(\epsilon/(2\mathfrak{L}); P_W, \varrho)/\log(2\mathfrak{L}/\epsilon) \leqslant 2\dim_{\mathrm{R}}(P_W).$$

Note that this holds due to the uniform convergence assumption of the corollary. Then, using Corollary 6, we have

$$\begin{aligned}|\mathbb{E}[\mathrm{gen}(S,W)]| \leqslant& \sqrt{\frac{2\sigma^2 \dim_{\mathrm{R}}(P_W,\delta)\log(n\mathfrak{L}^2)}{n}} + \sqrt{\frac{4}{n}} \\ \leqslant& \sqrt{\frac{4\sigma^2 \dim_{\mathrm{R}}(P_W,\delta)\log(n\mathfrak{L}^2)}{n}},\end{aligned}$$

where the last inequality holds for $n \geqslant n_0$, where $n_0 \geqslant n_0'$ is a sufficiently large integer. ■

E.5. Proof of Theorem 9

Proof Some of the steps in this proof are identical to the proof of Theorem 14, by considering X^m as $(\mathrm{gen}(S_1,W_1),\ldots,\mathrm{gen}(S_m,W_m))$ and $\hat{X}^m(j)$ as $(\mathrm{gen}(S_1,\hat{w}_{j,1}),\ldots,\mathrm{gen}(S_m,\hat{w}_{j,m}))$. Here, for the sake of completeness, we re-state all steps for the particular setup and notations used for the generalization error problem.

For any $\nu \in (0,\log(1/\delta))$ sufficiently small, choose m_0 such that for $m \geqslant m_0$,

$$-\frac{1}{m}\log\big(\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m(\mathcal{H}_m,\epsilon;\varphi_m))\big) \geqslant \log(1/\delta) - \nu. \tag{57}$$

For ease of notations, let $\mathcal{E}_m \coloneqq \mathcal{E}_m(\mathcal{H}_m, \epsilon; \varphi_m)$. Then,

$$
\begin{aligned}
&\mathbb{P}_{S,W}(\operatorname{gen}(S,W) \geqslant \Delta)^m \\
&\quad = \mathbb{P}_{(S,W)^{\otimes m}}(\forall i, \operatorname{gen}(S_i, W_i) \geqslant \Delta) \\
&\quad \leqslant \mathbb{P}_{(S,W)^{\otimes m}}(\forall i, \operatorname{gen}(S_i, W_i) \geqslant \Delta, \mathcal{E}_m^c) + \mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m) \\
&\quad \leqslant \mathbb{P}_{(S,W)^{\otimes m}}(\forall i, \operatorname{gen}(S_i, W_i) \geqslant \Delta, \mathcal{E}_m^c) + e^{-m(\log(1/\delta)-\nu)} \\
&\quad \leqslant \mathbb{P}_{S^{\otimes m}}(\exists w^m \colon \forall i, \operatorname{gen}(S_i, w_i) \geqslant \Delta, \mathcal{E}_m^c) + e^{-m(\log(1/\delta)-\nu)} \\
&\quad \leqslant \mathbb{P}_{S^{\otimes m}}\left(\exists j \in [l_m], \{\Delta_i\}_{i=1}^m \in \mathbb{R} \colon \forall i, \operatorname{gen}(S_i, \hat{w}_{j,i}) \geqslant \Delta - \Delta_i, \sum_{i=1}^m \Delta_i \leqslant m\epsilon\right) + e^{-m(\log(1/\delta)-\nu)} \\
&\quad \leqslant \mathbb{P}_{S^{\otimes m}}\left(\exists j \in [l_m], \{\Delta_i\}_{i=1}^m \in \mathbb{R} \colon \sum_{i=1}^m \operatorname{gen}(S_i, \hat{w}_{j,i}) \geqslant \sum_{i=1}^m (\Delta - \Delta_i), \sum_{i=1}^m \Delta_i \leqslant m\epsilon\right) + e^{-m(\log(1/\delta)-\nu)} \\
&\quad \leqslant \mathbb{P}_{S^{\otimes m}}\left(\exists j \in [l_m] \colon \sum_{i=1}^m \operatorname{gen}(S_i, \hat{w}_{j,i}) \geqslant m(\Delta - \epsilon)\right) + e^{-m(\log(1/\delta)-\nu)} \\
&\quad \leqslant \sum_{j \in [l_m]} \mathbb{P}_{S^{\otimes m}}\left(\sum_{i=1}^m \operatorname{gen}(S_i, \hat{w}_{j,i}) \geqslant m(\Delta - \epsilon)\right) + e^{-m(\log(1/\delta)-\nu)} \\
&\quad \overset{(a)}{\leqslant} \sum_{j \in [l_m]} e^{-mn(\Delta-\epsilon)^2/(2\sigma^2)} + e^{-m(\log(1/\delta)-\nu)} \\
&\quad \overset{(a)}{\leqslant} e^{m(R - n(\Delta-\epsilon)^2/(2\sigma^2))} + e^{-m(\log(1/\delta)-\nu)} \\
&\quad \overset{(b)}{\leqslant} 2e^{-m(\log(1/\delta)-\nu)}
\end{aligned}
$$

where (a) is derived using the Hoeffding's inequality, (b) is derived since $l_m \leqslant e^{mR}$, and (c) is derived by choosing Δ as $\Delta \coloneqq \sqrt{\frac{2\sigma^2(R+\log(1/\delta))}{n}} + \epsilon$. The proof completes by taking the m'th root of both sides, and since ν can be chosen arbitrarily small.

■

E.6. Proof of Theorem 10

Theorem 10 is stated in Section 3.3 for U being a constant and in Appendix D.2 has been extended to take into account the stochasticity of the algorithm. In the following, we first state the proof for finite sets and for the U being a constant using Theorem 9. The result can be extended to the case of arbitrary U that satisfies the conditions of the theorem, and to infinite sets, with some further assumptions on (S, W), using the quantization technique used in the proof of (El Gamal and Kim, 2011, Theorem 3.6) and by applying (Iriyama, 2005, Theorem 1) and its adaptation for the memoryless sources in (Bakshi and Bansal, 2005, Theorem 3). However, for the general case, we state an alternative proof that applies the Donsker–Varadhan's variational representation of the KL divergence.

E.6.1. FIRST PROOF

Proof Suppose that $\mathcal{S} \times \mathcal{W}$ is a finite set and U is a constant. We start by showing that for every $\epsilon \in \mathbb{R}$ and any $\nu_1, \nu_2 > 0$, the algorithm $\mathcal{A}(S)$ is $(R(\delta,\epsilon)+\nu_1, \epsilon+\nu_2, \delta; \{\varphi_m\}_m)$-exponentially compressible. Our proof is similar to (Marton, 1974, Theorem 1).

Let Q_m be an arbitrary type of $\mathcal{S}^m \times \mathcal{W}^m$. For the definition of the type, refer to the beginning of the appendices. Define

$$\mathcal{Q}_m(\delta) \coloneqq \{Q_m \colon D_{KL}(Q_m \| P_{S,W}) \leqslant \log(1/\delta)\}. \tag{58}$$

Note that,

$$\begin{aligned}
\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{T}(S^m,W^m) \notin \mathcal{Q}_m(\delta)) &= \sum_{Q'_m \notin \mathcal{Q}(\delta)} \mathbb{P}_{(S,W)^{\otimes m}}\big(\mathcal{T}(S^m,W^m) = Q'_m\big) \\
&\overset{(a)}{\leqslant} \sum_{Q'_m \notin \mathcal{Q}(\delta)} e^{-mD_{KL}\left(Q'_m \| P_{S,W}\right)} \\
&\overset{(b)}{\leqslant} m^{|\mathcal{S}|\times|\mathcal{W}|} e^{m\log(\delta)} \\
&= e^{m(\log(\delta)+|\mathcal{S}|\times|\mathcal{W}|\log(m)/m)} \\
&= (\delta+\varepsilon_m)^m,
\end{aligned} \tag{59}$$

where $\lim_{m\to\infty}\varepsilon_m = 0$, the step (a) is due to (Cover and Thomas, 2006, Theorem 11.1.4), and the step (b) is deduced since number of types can be bounded by $m^{|\mathcal{S}|\times|\mathcal{W}|}$.

First, we state a variant of type covering lemma (Berger, 1975, Section 6.1.2, Lemma 1) (appeared also in (Csiszár and Körner, 2011, Lemma 9.1)), proved in Appendix E.18:

Lemma 28 *For any $\nu > 0$ and any type Q_m, there exists a hypothesis book $\mathcal{H}_{Q_m} = \{\hat{\mathbf{w}}_{Q_m,j}, j \in [l_{Q_m}]\} \subseteq \hat{\mathcal{W}}^m$, such that $l_{Q_m} \leqslant e^{mR_{Q_m}}$, where*

$$R_{Q_m} = \mathfrak{RD}^*(\epsilon; Q_m) + \varepsilon_{Q_m}, \tag{60}$$

and $\lim_{m\to\infty}\varepsilon_{Q_m} = 0$, and such that for $m \geqslant m_{\nu,Q_m}$,

$$\mathbb{P}_{(S,W)^{\otimes m}}\left(\mathcal{T}(S^m,W^m) = Q_m, \min_{j\in[l_{Q_m}]} \varphi_m(W^m, \hat{\mathbf{w}}_{Q_m,j}; S^m) \geqslant \epsilon+\nu\right) = 0. \tag{61}$$

Let $m \geqslant m_\nu$, where m_ν is sufficiently large such that the above lemma holds for all types. Letting $\mathcal{H}_m \coloneqq \bigcup_{Q_m\in\mathcal{Q}_m(\delta)} \mathcal{H}_{Q_m}$, we have

$$\begin{aligned}
l_m &\leqslant \sum_{Q_m\in\mathcal{Q}_m(\delta)} l_{Q_m} \\
&\leqslant m^{|\mathcal{S}|\times|\mathcal{W}|} e^{m\left(\max_{Q_m\in\mathcal{Q}_m(\delta)}\left[\mathfrak{RD}^*(\epsilon;Q_m)+\varepsilon_{Q_m}\right]\right)} \\
&= e^{m\left(\max_{Q_m\in\mathcal{Q}_m(\delta)}\mathfrak{RD}^*(\epsilon;Q_m)+\varepsilon'_m\right)},
\end{aligned} \tag{62}$$

where $\lim_{m\to\infty}\varepsilon'_m = 0$. Moreover,

$$\begin{aligned}
&\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m(\mathcal{H}_m, \epsilon+\nu;\varphi_m))\\
&\quad\leqslant \mathbb{P}_{(S,W)^{\otimes m}}((S^m,W^m)\notin \mathcal{Q}_m)\\
&\qquad + \mathbb{P}_{(S,W)^{\otimes m}}\left((S^m,W^m)\in\mathcal{Q}_m, \min_{Q_m\in\mathcal{Q}_m}\min_{j\in[l_{Q_m}]}\varphi_m(W^m,\hat{\mathbf{w}}_{Q_m,j};S^m)\geqslant \epsilon+\nu\right)\\
&\quad\overset{(a)}{\leqslant}(\delta+\varepsilon_m)^m.
\end{aligned}$$

where the last steps is derived using Lemma 28 and relation (59).

Hence,

$$\lim_{m\to\infty} -\frac{1}{m}\log\big(\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m(\mathcal{H}_m,\epsilon+\nu;\varphi_m))\big)\leqslant \log(1/\delta). \tag{63}$$

The relations (62) and (63) show that for every $\epsilon\in\mathbb{R}$ and any $\nu_1,\nu_2>0$, the algorithm $\mathcal{A}(S)$ is $(R(\delta,\epsilon)+\nu_1,\epsilon+\nu_2,\delta;\{\varphi_m\}_m)$-exponentially compressible. Using Theorem 9 completes the proof. ■

E.6.2. SECOND PROOF

Proof In this proof, since we use Theorem 12, we denote $P_{U,S,W}$ by $\mu_{U,S,W}$, to be compatible with the notations of Theorem 12. Note that $\mu_{U,S,W}=\mu_S\mu_U\mu_{W|S,U}$ where U is an independent noise in the algorithm and W is the output hypothesis of the algorithm.

Let $X\coloneqq \mathrm{gen}(W,S)$. This defines $\mu_{S,U,W,X}$. Writing Theorem 12 with the choice of $Y=(S,U,W)$, we get that

$$\begin{aligned}
&\log\mathbb{P}_{X\sim\mu_X}(X\geqslant\Delta) \qquad\qquad (64)\\
&\leqslant \max\Bigg[\log(\delta), \sup_{\nu_{S,U,W,X}\in\mathcal{G}}\ \inf_{p_{\hat{X}|S,U,W}\in\mathcal{Q}(\nu),\, q_{\hat{X}|S,U,W},\,\lambda\geqslant 0}\Big\{D_{KL}(p_{\hat{X}|S,U,W}\nu_{S,U,W}\|q_{\hat{X}|S,U,W}\nu_{S,U,W})\\
&\qquad\qquad -\lambda(\Delta-\epsilon)+\log\mathbb{E}_{\mu_{S,U,W}q_{\hat{X}|S,U,W}}[e^{\lambda\hat{X}}]\Big\}\Bigg]
\end{aligned}$$

where $\mathcal{G}$ is the following set of distributions:

$$\mathcal{G}=\{\nu_{X,S,U,W}: D_{KL}(\nu_{X,S,U,W}\|\mu_{X,S,U,W})\leqslant\log(1/\delta),\}$$

and $\mathcal{Q}(\nu)$ is the set of conditional distributions $p_{\hat{X}|S,U,W}$ such that under $p_{\hat{X}|S,U,W}\nu_{S,U,W}$ we have:

$$\left[\inf_{x\in\mathrm{supp}(\nu_X)} x\right]-\mathbb{E}[\hat{X}]\leqslant\epsilon.$$

Since $X=\mathrm{gen}(W,S)$ under $\mu_{X,U,W,S}$, and $D_{KL}(\nu_{X,S,U,W}\|\mu_{X,S,U,W})<\infty$, we obtain that $X=\mathrm{gen}(W,S)$ under $\nu_{X,U,W,S}$ too. Therefore, the supremum is over distribution ν of the form $\nu_{X,U,W,S}=\nu_{U,W,S}\mu_{X|U,W,S}$. This implies that

$$D_{KL}(\nu_{X,S,U,W}\|\mu_{X,S,U,W})=D_{KL}(\nu_{S,U,W}\|\mu_{S,U,W})$$

and

$$\inf_{x\in\mathrm{supp}(\nu_X)} x = \inf_{(s',w')\in\mathrm{supp}(\nu_{S,W})} \big[\mathrm{gen}(s',w')\big].$$

Now, take some arbitrary $\nu_{S,U,W}$, and also take some arbitrary $p_{\hat{W}_1|U,S,W}$ satisfying

$$p_{\hat{W}_1|U,S,W} = p_{\hat{W}_1|U,S}$$

where $\hat{W}_1 \in \mathcal{W}$ belongs to the hypothesis space and

$$\inf_{(s',w')\in\mathrm{supp}(\nu_{S,W})} \big[\mathrm{gen}(s',w')\big] - \mathbb{E}_{\nu_{U,S}p_{\hat{W}_1|U,S}}\Big[\mathrm{gen}(\hat{W}_1,S)\Big] \leqslant \epsilon.$$

Let $\hat{X}_1 = \mathrm{gen}(\hat{W}_1, S)$. Then, $p_{\hat{X}_1|S,U,W} \in \mathcal{Q}(\nu)$.

Next, we define $\hat{W}_2$ to have a joint distribution of the form $p_{\hat{W}_2|U}\nu_{S,U,W,X}$ such that the marginal joint distribution of $(\hat{W}_2, U)$ is the same as $(\hat{W}_1, U)$ under $p_{\hat{W}_1|U,S}\nu_{S,U,W,X}$. Note that $p_{\hat{W}_2|U,S} = p_{\hat{W}_2|U}$ is assumed here by the fact that $\hat{W}_2$ has a joint distribution of the form $p_{\hat{W}_2|U}\nu_{S,U,W,X}$.

Let $\hat{X}_2 = \mathrm{gen}(\hat{W}_2, S)$. Take $q_{\hat{X}_2|S,U,W}$ to be the conditional distribution of $\hat{X}_2$ given S, U, W. Also, take $p_{\hat{X}_1|S,U,W} \in \mathcal{Q}(\nu)$ to be the conditional distribution of $\hat{X}_1$ given S, U, W. We evaluate the above bound with $q_{\hat{X}_2|S,U,W}$ and $p_{\hat{X}_1|S,U,W}$. Then,

$$\begin{aligned} D_{KL}(p_{\hat{X}_2|S,U,W}\nu_{S,U,W}\|q_{\hat{X}_1|S,U,W}\nu_{S,U,W}) &\leqslant D_{KL}(p_{\hat{W}_1|S,U}\nu_{S,U,W}\|p_{\hat{W}_2|S,U}\nu_{S,U,W}) \\ &= D_{KL}(p_{\hat{W}_1|S,U}\nu_{S,U,W}\|p_{\hat{W}_1|U}\nu_{S,U,W}) \\ &= D_{KL}(p_{\hat{W}_1|S,U}\nu_{S,U}\|p_{\hat{W}_1|U}\nu_{S,U}) \\ &= I_{p_{\hat{W}_1|S,U}\nu_{S,U}}(\hat{W}_1; S|U). \end{aligned}$$

Finally, under $\mu_{U,S,W}q_{\hat{X}_2|S,U,W}$ we have that $S = (Z_1, \ldots, Z_n)$ is an i.i.d. sequence according to μ_Z. Moreover, in $\mu_{U,S}$ we have that U is independent of S. Furthermore, in $q_{\hat{X}_2|S}$, we have $p_{\hat{W}_2|U,S} = p_{\hat{W}_2|U}$, which together with independence of U and S implies that $\hat{W}_2$ is independent of S. Therefore, $\mathrm{gen}(\hat{W}_2, S)$ is the sum of n i.i.d. variables, and since $\ell(Z, \hat{w})$ is σ-subgaussian, hence $\mathrm{gen}(\hat{W}_2, S)$ is $\sigma/\sqrt{n}$-subgaussian. Therefore, we can compute its moment generating function. Thus, $\log \mathbb{E}\Big[e^{\lambda\hat{X}}\Big] \leqslant \lambda^2\sigma^2/2n$ and letting $\lambda \coloneqq n(\Delta-\epsilon)/\sigma^2$ yields

$$\begin{aligned} D_{KL}(p_{\hat{X}|S,U}\nu_{S,U}\|q_{\hat{X}|S,U}\nu_{S,U}) - \lambda(\Delta-\epsilon) + \log \mathbb{E}_{\mu_{S,U}q_{\hat{X}|S,U}}[e^{\lambda\hat{X}}] & \\ \leqslant I_{p_{\hat{W}_1|S,U}\nu_{S,U}}(\hat{W}_1; S|U) - \frac{n(\Delta-\epsilon)^2}{2\sigma^2}. & \end{aligned}$$

Hence, denoting $\mathcal{G}_{S,U,W} = \{\nu_{S,U,W} : D_{KL}(\nu_{S,U,W}\|\mu_{S,U,W}) \leqslant \log(1/\delta)\}$, and by definition (14), we have

$$\begin{aligned} &\log \mathbb{P}_{X\sim\mu_X}(X \geqslant \Delta) \\ &\leqslant \max\left[\log(\delta), \sup_{\nu_{S,U,W}\in\mathcal{G}_{S,U,W}} \left\{\mathfrak{RD}^*(\epsilon; Q) - \frac{n(\Delta-\epsilon)^2}{2\sigma^2}\right\}\right]. \end{aligned} \tag{65}$$

Letting

$$\Delta = \sup_{\nu_{S,U,W} \in \mathcal{G}_{S,U,W}} \sqrt{\frac{2\sigma^2(\mathfrak{RD}^*(\epsilon;Q) + \log(1/\delta))}{n}} + \epsilon,$$

completes the proof. ■

E.7. Proof of Theorem 12

We first claim that

$$\log \mathbb{P}_{X\sim\mu_X}(X \geqslant \Delta)$$
$$\leqslant \sup_{\nu_{YX} \ll \mu_{YX}} \inf_{p_{\hat{X}|Y} \in \mathcal{Q}(\nu),\, \lambda \geqslant 0} \left\{ -D_{KL}(\nu_{YX}\|\mu_{YX}) - \lambda \left[(\Delta - \epsilon) - \int \hat{x} d(p_{\hat{X}|Y}\nu_{Y,X}) \right]_+ \right\}. \quad (66)$$

This follows from Lemma 24 because if we look at ν_{YX}'s of the form $\nu_{YX} = \nu_X \mu_{Y|X}$, we have

$$D_{KL}(\nu_{YX}\|\mu_{YX}) = D_{KL}(\nu_X\|\mu_X).$$

Moreover, the term $\int \hat{x} d(p_{\hat{X}|Y}\nu_{Y,X})$ depends only on the marginal distribution on $\hat{X}$ under $p_{\hat{X}|Y}\nu_{Y,X}$.

Next, given $p_{\hat{X}|Y}$, let $p_{\hat{X},Y} = p_{\hat{X}|Y}\nu_Y$. For every $Y = y$, consider the conditional distribution $p_{\hat{X}|Y=y}$ induced by this joint distribution. Lemma 23 yields

$$\inf_{q_{\hat{X}}} \left[D_{KL}(p_{\hat{X}|Y=y}\|q_{\hat{X}}) + \log \mathbb{E}_q[e^{\lambda \hat{X}}] \right] = \lambda \mathbb{E}_p[\hat{X}|Y = y], \quad (67)$$

By averaging this over y using the distribution ν_Y, we obtain

$$\inf_{q_{\hat{X}|Y}} \left[D_{KL}(p_{\hat{X}|Y}\nu_Y\|q_{\hat{X}|Y}\nu_Y) + \mathbb{E}_{\nu_Y} \log \mathbb{E}_q[e^{\lambda \hat{X}}|Y] \right] = \lambda \int \hat{x} d(p_{\hat{X}|Y}\nu_Y). \quad (68)$$

This equality along with (66) yield

$$\log \mathbb{P}_{X\sim\mu_X}(X \geqslant \Delta)$$
$$\leqslant \sup_{\nu_{YX} \ll \mu_{YX}} \inf_{p_{\hat{X}|Y} \in \mathcal{Q}(\nu)} \inf_{q_{\hat{X}|Y},\, \lambda \geqslant 0} \Bigg\{ - D_{KL}(\nu_{YX}\|\mu_{YX})$$
$$- \left[\lambda(\Delta - \epsilon) - D_{KL}(p_{\hat{X}|Y}\nu_Y\|q_{\hat{X}|Y}\nu_Y) - \mathbb{E}_{Y\sim\nu_Y} \log \mathbb{E}_q[e^{\lambda \hat{X}}|Y] \right]_+ \Bigg\}.$$

From the Donsker-Varadhan's identity we obtain the inequality

$$\mathbb{E}_{Y\sim\nu_Y} \log \mathbb{E}_q[e^{\lambda \hat{X}}|Y] \leqslant D_{KL}(\nu_Y\|\mu_Y) + \log \mathbb{E}_{\mu_Y} \mathbb{E}_q[e^{\lambda \hat{X}}|Y] = D_{KL}(\nu_Y\|\mu_Y) + \log \mathbb{E}_{\mu_Y q_{\hat{X}|Y}} \mathbb{E}[e^{\lambda \hat{X}}].$$

Therefore,

$$\log \mathbb{P}_{X\sim\mu_X}(X \geqslant \Delta)$$
$$\leqslant \sup_{\nu_{YX} \ll \mu_{YX}} \inf_{p_{\hat{X}|Y} \in \mathcal{Q}(\nu),\, q_{\hat{X}|Y},\, \lambda \geqslant 0} \Bigg\{ - D_{KL}(\nu_{YX}\|\mu_{YX}) - \Bigg[\lambda(\Delta - \epsilon) - D_{KL}(p_{\hat{X}|Y}\nu_Y\|q_{\hat{X}|Y}\nu_Y)$$
$$- D_{KL}(\nu_Y\|\mu_Y) - \log \mathbb{E}_{\mu_Y q_{\hat{X}|Y}} \mathbb{E}[e^{\lambda \hat{X}}] \Bigg]_+ \Bigg\}.$$

The desired inequality follows from here since $D_{KL}(\nu_{YX}\|\mu_{YX}) \geqslant D_{KL}(\nu_Y\|\mu_Y)$.

E.8. Proof of Theorem 14

Proof Let $X^m = (X_1, \dots, X_m)$. We can write

$$
\begin{aligned}
&\mathbb{P}(X \geqslant \Delta)^m \\
=&\mathbb{P}\left(\min_{i\in[m]} X_i \geqslant \Delta\right) \\
\leqslant&\mathbb{P}\left(\min_{i\in[m]} X_i \geqslant \Delta, \min_{j\in[k]} \rho(X^m, \hat{X}^m(j)) \leqslant \epsilon\right) + \mathbb{P}\left(\min_{j\in[k]} \rho(X^m, \hat{X}^m(j)) > \epsilon\right) \\
=&\mathbb{P}\left(\exists j \in [k] : \min_{i\in[m]} X_i \geqslant \Delta, \rho(X^m, \hat{X}^m(j)) \leqslant \epsilon\right) + \mathbb{P}\left(\min_{j\in[k]} \rho(X^m, \hat{X}^m(j)) > \epsilon\right) \\
=&\mathbb{P}\left(\exists j \in [k] : \rho(X^m, \hat{X}^m(j)) + \frac{1}{m}\sum_{i=1}^m \hat{X}_i(j) \geqslant \Delta, \rho(X^m, \hat{X}^m(j)) \leqslant \epsilon\right) \qquad (69) \\
&\quad + \mathbb{P}\left(\min_{j\in[k]} \rho(X^m, \hat{X}^m(j)) > \epsilon\right) \\
\leqslant&\mathbb{P}\left(\exists j \in [k] : \frac{1}{m}\sum_{i=1}^m \hat{X}_i(j) \geqslant \Delta - \epsilon\right) + \mathbb{P}\left(\min_{j\in[k]} \rho(X^m, \hat{X}^m(j)) > \epsilon\right) \\
\leqslant&\sum_{j=1}^k \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \hat{X}_i(j) \geqslant \Delta - \epsilon\right) + \mathbb{P}\left(\min_{j\in[k]} \rho(X^m, \hat{X}^m(j)) > \epsilon\right),
\end{aligned}
$$

where (69) follows from the definition of ρ (Definition 13). ■

E.9. Proof of Theorem 16

Proof First, note that as established in (Steinke and Zakynthinou, 2020, Proof of Theorem 5.1),

$$\mathbb{E}[\mathrm{gen}(S, W)] = \mathbb{E}_{\mathfrak{Z}}\mathbb{E}_{\mathbf{K},W|\mathfrak{Z}}[f(\mathfrak{Z}, \mathbf{K}, W)].$$

The rest of the proof is similar to the proof of Theorem 2, by considering the term $f(\mathfrak{Z}, \mathbf{K}, W)$, instead of $\mathrm{gen}(S, W)$, and by noting that conditioned on $\mathfrak{Z} = \mathfrak{z}$ and $W = w$, for every $j \in [n]$, $(-1)^{K_j}(\ell(\mathfrak{z}_{j,1}, w) - \ell(\mathfrak{z}_{j,2}, w))$ is a bounded process in the range $[-1, 1]$, with average zero, that takes values among $\ell(\mathfrak{z}_{j,1}, w) - \ell(\mathfrak{z}_{j,2}, w)$ and $-(\ell(\mathfrak{z}_{j,1}, w) - \ell(\mathfrak{z}_{j,2}, w))$, uniformly. Hence, $f(\mathfrak{z}, \mathbf{K}, w)$ is $1/\sqrt{n}$-subgaussian. We show the proof for part i. The other parts follow similarly. Let $\hat{\mathbf{w}}(\mathfrak{z}, \mathbf{k}^m, w^m) \coloneqq \hat{\mathbf{w}}_j(\mathfrak{z})$ where $j = \underset{j\in[l_m(\mathfrak{z})]}{\arg\min}\, \vartheta_m(w^m, \hat{\mathbf{w}}_j(\mathfrak{z}); \mathfrak{z}, \mathbf{k}^m)$. We denote it simply by $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_m)$. Then, we have

$$
\begin{aligned}
&\mathbb{E}_{(\mathbf{K},W|\mathfrak{z})}[f(\mathfrak{z}, \mathbf{K}, W)] \\
&\quad\leqslant \frac{1}{m}\mathbb{E}_{(\mathbf{K},W|\mathfrak{z})^{\otimes m}}\left[\sum_{i=1}^m f(\mathfrak{z}, \mathbf{K}_i, W_i)\right] \\
&\quad\leqslant \frac{1}{m}\mathbb{E}_{(\mathbf{K},W|\mathfrak{z})^{\otimes m}}\left[\sum_{i=1}^m f(\mathfrak{z}, \mathbf{K}_i, W_i) - f(\mathfrak{z}, \mathbf{K}_i, \hat{W}_i)\right] + \frac{1}{m}\mathbb{E}_{(\mathbf{K},W|\mathfrak{z})^{\otimes m}}\left[\sum_{i=1}^m f(\mathfrak{z}, \mathbf{K}_i, \hat{W}_i)\right]
\end{aligned}
$$

$$\leqslant \frac{1}{m}\mathbb{E}_{\mathbf{K}^{\otimes m}}\left[\max_{j\in[l_m(\mathfrak{z})]}\sum_{i=1}^{m} f(\mathfrak{z},\mathbf{K}_i,\hat{w}_{j,i})\right] + \epsilon(\mathfrak{z}) + \varepsilon_m$$
$$\leqslant \frac{1}{m}\sqrt{\frac{2m\log(l_m(\mathfrak{z}))}{n}} + \epsilon(\mathfrak{z}) + \varepsilon_m$$
$$\leqslant \sqrt{\frac{2R(\mathfrak{z})}{n}} + \epsilon(\mathfrak{z}) + \varepsilon_m.$$

Taking the limit for $m \to \infty$ completes the proof. ■

E.10. Proof of Corollary 18

Proof Let U be the stochasticity of the algorithm (*e.g.* the randomness in choosing the training data for each batch in the SGD algorithm) in a sense that for a given dataset $s = (z_1, \ldots, z_n)$ and based on the sequence of values $\{(\ell(z_1, w), \ldots, \ell(z_n, w))\}_{w\in\mathcal{W}}$, the algorithm chooses a fixed hypothesis w conditioned on $U = u$.

If an algorithm has the VC-dimension d and for a fixed $\mathfrak{Z} = \mathfrak{z}$, the set of possible pairs $\{(\ell(z_1, w), \ldots, \ell(z_n, w))\}_w$, where $z_i = \mathfrak{z}_{\mathbf{k}}$ for some $\mathbf{k} \in \{1, 2\}^n$, is bounded by the set of possible $\{(\ell(z_{1,1}, w), \ell(z_{1,2}, w), \ldots, \ell(z_{n,1}, w), \ell(z_{n,2}, w))\}_w$, and the latter is bounded by $(2en/d)^d$ due to Sauer-Shelah lemma Sauer (1972); Shelah (1972). Hence, $I(\mathbf{K}, W|\mathfrak{z}, u) \leqslant d\log(2en/d)$. Using Theorem 17 completes the proof. ■

E.11. Proof of Theorem 19

Proof Let $\Delta_1 \coloneqq \sqrt{\log(2/\delta)/n}$, $\Delta_2 \coloneqq \sup_{\mathfrak{z}} \sqrt{2(R(\mathfrak{z}) + \log(2/\delta))/n} + \epsilon(\mathfrak{z})$, and let $\bar{S} \sim \mu^{\otimes n}$ be independent of (S, W). Then,

$$\begin{aligned}
\mathbb{P}(\mathrm{gen}(S, W) \geqslant \Delta_1 + \Delta_2) =& \mathbb{P}\Big(\mathbb{E}_{\bar{S}}\Big[\hat{\mathcal{L}}(\bar{S}, W)\Big] - \hat{\mathcal{L}}(\bar{S}, W) + \hat{\mathcal{L}}(\bar{S}, W) - \hat{\mathcal{L}}(S, W) \geqslant \Delta_1 + \Delta_2\Big) \\
\leqslant& \mathbb{P}\Big(\mathbb{E}_{\bar{S}}\Big[\hat{\mathcal{L}}(\bar{S}, W)\Big] - \hat{\mathcal{L}}(\bar{S}, W) \geqslant \Delta_1\Big) + \mathbb{P}\Big(\hat{\mathcal{L}}(\bar{S}, W) - \hat{\mathcal{L}}(S, W) \geqslant \Delta_2\Big) \\
\leqslant& \delta/2 + \mathbb{P}\Big(\hat{\mathcal{L}}(\bar{S}, W) - \hat{\mathcal{L}}(S, W) \geqslant \Delta_2\Big).
\end{aligned}$$

It remains to upper bound the second term by $\delta/2$. Denote $\mathfrak{Z} \in \mathcal{Z}^{2\times n}$ as concatenation of S and $\bar{S}$, such that for some $\mathbf{K} \in \{1, 2\}^n$, $Z_i = \mathfrak{Z}_{i,K_i}$ and $\bar{Z}_i = \mathfrak{Z}_{i,\bar{K}_i}$. As before, we denote $S = \mathfrak{Z}_{\mathbf{K}}$ and $\bar{S} = \mathfrak{Z}_{\bar{K}}$. The joint distribution of $\mathfrak{Z}, \mathbf{K}, W$ is $P_{\mathfrak{Z}}P_{\mathbf{K}}P_{W|\mathfrak{Z}_{\mathbf{K}}}$, where $P_{\mathfrak{Z}} = \mu^{\otimes 2n}$ and $P_{\mathbf{K}}$ is uniform over $\{1, 2\}^n$. Now,

$$\begin{aligned}
\mathbb{P}\Big(\hat{\mathcal{L}}(\bar{S}, W) - \hat{\mathcal{L}}(S, W) \geqslant \Delta_2\Big) =& \mathbb{P}\Big(\hat{\mathcal{L}}(\mathfrak{Z}_{\bar{\mathbf{K}}}, W) - \hat{\mathcal{L}}(\mathfrak{Z}_{\mathbf{K}}, W) \geqslant \Delta_2\Big) \\
=& \mathbb{P}(f(\mathfrak{Z}, \mathbf{K}, W) \geqslant \Delta_2) \\
\leqslant& \max_{\mathfrak{z}} \mathbb{P}(f(\mathfrak{z}, \mathbf{K}, W) \geqslant \Delta_2).
\end{aligned}$$

The rest of proof is to bound $\mathbb{P}_{\mathbf{K},W|\mathfrak{z}}(f(\mathfrak{z}, \mathbf{K}, W) \geqslant \Delta_2)$ for a fixed $\mathfrak{z}$. Similar to the proof of Theorem 9, for any $\nu \in (0, \log(2/\delta))$ sufficiently small, choose m_0 such that for $m \geqslant m_0$,

$$\lim_{m\to\infty}\left[-\frac{1}{m}\log\left(\mathbb{P}_{(\mathbf{K},W|\mathfrak{z})^{\otimes m}}\left(\min_{j\in[l_m(\mathfrak{z})]}\varphi_m(W^m,\hat{\mathbf{w}}_j(\mathfrak{z});\mathfrak{z},\mathbf{K}^m)>\epsilon(\mathfrak{z})\right)\right)\right]\geqslant\log(2/\delta)-\nu.$$

For ease of notations, let $\mathcal{E}_m$ be the event that $\min_{j\in[l_m(\mathfrak{z})]}\varphi_m(W^m,\hat{\mathbf{w}}_j(\mathfrak{z});\mathfrak{z},\mathbf{K}^m)>\epsilon(\mathfrak{z})$. Then,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{K},W|\mathfrak{z}}(f(\mathfrak{z},\mathbf{K},W)\geqslant\Delta_2)^m\\
=&\mathbb{P}_{(\mathbf{K},W|\mathfrak{z})^{\otimes m}}(\forall i,f(\mathfrak{z},\mathbf{K}_i,W_i)\geqslant\Delta_2)\\
\leqslant&\mathbb{P}_{(\mathbf{K},W|\mathfrak{z})^{\otimes m}}(\forall i,f(\mathfrak{z},\mathbf{K}_i,W_i)\geqslant\Delta_2,\mathcal{E}_m^c)+\mathbb{P}_{(\mathbf{K},W|\mathfrak{z})^{\otimes m}}(\mathcal{E}_m)\\
\leqslant&\mathbb{P}_{(\mathbf{K},W|\mathfrak{z})^{\otimes m}}(\forall i,f(\mathfrak{z},\mathbf{K}_i,W_i)\geqslant\Delta_2,\mathcal{E}_m^c)+e^{-m(\log(2/\delta)-\nu)}\\
\leqslant&\mathbb{P}_{\mathbf{K}^{\otimes m}}(\exists w^m\colon\forall i,f(\mathfrak{z},\mathbf{K}_i,w_i)\geqslant\Delta_2,\mathcal{E}_m^c)+e^{-m(\log(2/\delta)-\nu)}\\
\leqslant&\mathbb{P}_{\mathbf{K}^{\otimes m}}\left(\exists j\in[l_m(\mathfrak{z})],\{\Delta_i\}_{i=1}^m\in\mathbb{R}\colon\forall i,f(\mathfrak{z},\mathbf{K}_i,\hat{w}_{j,i})\geqslant\Delta_2-\Delta_i,\sum_{i=1}^m\Delta_i\leqslant m\epsilon(\mathfrak{z})\right)+e^{-m(\log(2/\delta)-\nu)}\\
\leqslant&\mathbb{P}_{\mathbf{K}^{\otimes m}}\left(\exists j\in[l_m(\mathfrak{z})],\{\Delta_i\}_{i=1}^m\in\mathbb{R}\colon\sum_{i=1}^m f(\mathfrak{z},\mathbf{K}_i,\hat{w}_{j,i})\geqslant\sum_{i=1}^m(\Delta_2-\Delta_i),\sum_{i=1}^m\Delta_i\leqslant m\epsilon(\mathfrak{z})\right)+e^{-m(\log(2/\delta)-\nu)}\\
\leqslant&\mathbb{P}_{\mathbf{K}^{\otimes m}}\left(\exists j\in[l_m(\mathfrak{z})]\colon\sum_{i=1}^m f(\mathfrak{z},\mathbf{K}_i,\hat{w}_{j,i})\geqslant m(\Delta_2-\epsilon(\mathfrak{z}))\right)+e^{-m(\log(2/\delta)-\nu)}\\
\leqslant&\sum_{j\in[l_m(\mathfrak{z})]}\mathbb{P}_{\mathbf{K}^{\otimes m}}\left(\sum_{i=1}^m f(\mathfrak{z},\mathbf{K}_i,\hat{w}_{j,i})\geqslant m(\Delta_2-\epsilon(\mathfrak{z}))\right)+e^{-m(\log(2/\delta)-\nu)}\\
\overset{(a)}{\leqslant}&\sum_{j\in[l_m(\mathfrak{z})]}e^{-mn(\Delta_2-\epsilon(\mathfrak{z}))^2/2}+e^{-m(\log(2/\delta)-\nu)}\\
\overset{(b)}{\leqslant}&e^{m(R(\mathfrak{z})-n(\Delta_2-\epsilon(\mathfrak{z}))^2/2)}+e^{-m(\log(2/\delta)-\nu)}\\
\overset{(c)}{\leqslant}&2e^{-m(\log(2/\delta)-\nu)},
\end{aligned}$$

where (a) is derived using the Hoeffding's inequality, (b) is derived since $l_m(\mathfrak{z})\leqslant e^{mR(\mathfrak{z})}$, and (c) is derived since $\Delta_2\geqslant\sqrt{\frac{2(R(\mathfrak{z})+\log(2/\delta))}{n}}+\epsilon(\mathfrak{z})$. The proof completes by taking the m'th root of both sides, and since ν can be chosen arbitrarily small. ■

E.12. Proof of Theorem 20

Proof First, similar to the proof of Theorem 19, we have

$$\mathbb{P}(\mathrm{gen}(S,W)\geqslant\Delta_1+\Delta_2)\leqslant\delta/2+\max_{\mathfrak{z}}\mathbb{P}(f(\mathfrak{z},\mathbf{K},W)\geqslant\Delta_2).$$

where $\Delta_1\coloneqq\sqrt{\log(2/\delta)/n}$, $\Delta_2\coloneqq\sup_{\mathfrak{z}}\sqrt{2(R(\mathfrak{z},\delta,\epsilon)+\log(2/\delta))/n}+\epsilon(\mathfrak{z})$, and $R(\mathfrak{z},\delta,\epsilon)$ is defined in the theorem. The rest of the proof is to upper bound $\mathbb{P}_{\mathbf{K},W|\mathfrak{z}}(f(\mathfrak{z},\mathbf{K},W)\geqslant\Delta_2)$ by $\delta/2$ for a fixed $\mathfrak{z}$, which follows similarly as the proof of Theorem 10, by considering $f(\mathfrak{z},\mathbf{K},W)$ instead of $\mathrm{gen}(S,W)$. ■

E.13. Proof of Corollary 22

Proof For any $\mathfrak{z}$ and under any Q such that $D_{KL}(Q\|P_{\mathbf{K},W,U|\mathfrak{z}}) \leqslant \log(2/\delta) < \infty$, we have $I(\mathbf{K}, W|u) \leqslant d\log(2en/d)$. Since, for any fixed U, similar to the proof of Corollary 18, the set of possible W under $P_{\mathbf{K},W,U|\mathfrak{z}}$ is bounded by $(2en/d)^d$, and consequently under Q as well. Using Theorem 20 completes the proof. ■

E.14. Proof of Lemma 23

Proof The Donsker-Varadhan's identity states that

$$D_{KL}(\nu\|\mu) = \sup_{\Phi}\Big\{\mathbb{E}_\nu[\Phi(X)] - \log\mathbb{E}_\mu[e^{\Phi(X)}]\Big\}.$$

The choice of $\Phi(\hat{x}) = \lambda\hat{x}$ implies the following inequality for any distributions ν and μ:

$$D_{KL}(\nu\|\mu) \geqslant \lambda\mathbb{E}_\nu[X] - \log\mathbb{E}_\mu[e^{\lambda X}].$$

Therefore,

$$\inf_{\mu}\Big[D_{KL}(\nu\|\mu) + \log\mathbb{E}_\mu[e^{\lambda X}]\Big] \geqslant \lambda\mathbb{E}_\nu[X].$$

On the other hand, if $\frac{d\mu}{d\nu}$ is proportional to $e^{-\lambda x}$, one can directly verify that

$$D_{KL}(\nu\|\mu) + \log\mathbb{E}_\mu[e^{\lambda X}] = \lambda\mathbb{E}_\nu[X].$$

Thus, the desired inequality is established.

■

E.15. Proof of Lemma 24

We state two proofs for this lemma.

E.15.1. FIRST PROOF

Proof We simplify the right hand side and reduce it to the left hand side. For any ν_X, if there exists a distribution $p_{\hat{X}} \in \mathcal{P}(\nu_X)$ such that $(\Delta - \epsilon) > \mathbb{E}_p[\hat{X}]$, then one can set $\lambda = \infty$. Otherwise, it is optimal to set $\lambda = 0$. Let $\mathcal{A}$ denote the set of distributions ν_X such that for any $p_{\hat{X}} \in \mathcal{P}(\nu_X)$ we have $(\Delta - \epsilon) \leqslant \mathbb{E}_p[\hat{X}]$. Then, the desired equality is equivalent with

$$\log\mathbb{P}_{X\sim\mu_X}(X \geqslant \Delta) = \sup_{\nu_X\in\mathcal{A}} -D_{KL}(\nu_X\|\mu_X).$$

Remember that $\mathcal{P}(\nu_X)$ denotes the set of distributions $p_{\hat{X}}$ on $\mathbb{R}$ for which

$$\Big[\inf_{x\in\text{supp}(\nu_X)} x\Big] - \mathbb{E}\Big[\hat{X}\Big] \leqslant \epsilon.$$

Thus, $\mathcal{A}$ is the set of distributions ν_X such that $\Delta \leqslant \inf_{x\in\text{supp}(\nu_X)} x$, or equivalently, $\text{supp}(\nu_X) \subseteq [\Delta, \infty)$. The minimum of $D_{KL}(\nu_X\|\mu_X)$ is then obtained by a distribution that is proportional with μ_X on $[\Delta, \infty)$, and the minimum value of $D_{KL}(\nu_X\|\mu_X)$ equals $-\log\mathbb{P}_{X\sim\mu_X}(X \geqslant \Delta)$. This completes the proof. ■

E.15.2. SECOND PROOF

Proof In this part, we give a second proof of Lemma 24 from Appendix C. Fix some natural number m. The *type* of a given sequence in $\mathcal{X}^m$ is defined as its empirical distribution. For every type ν_X of the sequences in $\mathcal{X}^m$ of length m, pick an arbitrary *type* $p_\nu(\hat{x})$ on a set $\hat{\mathcal{X}}$ satisfying

$$\left[\min_{x\colon \nu(x)>0} x\right] - \mathbb{E}_{p_\nu}\left[\hat{X}\right] \leqslant \epsilon. \tag{70}$$

Let $q_\nu(\hat{x})$ be another distribution such that

$$q_\nu(\hat{x}) = \frac{p_\nu(\hat{x})e^{-\lambda\hat{x}}}{\mathbb{E}_{p_\nu}e^{-\lambda\hat{X}}}, \qquad \forall \hat{x}.$$

Equivalently,

$$p_\nu(\hat{x}) = \frac{q_\nu(\hat{x})e^{\lambda\hat{x}}}{\mathbb{E}_{q_\nu}e^{\lambda\hat{X}}}, \qquad \forall \hat{x}.$$

Then, one can directly verify that

$$D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})) = \lambda\mathbb{E}_{p_\nu}[\hat{X}] - \log \mathbb{E}_{q_\nu}[e^{\lambda\hat{X}}]. \tag{71}$$

Let $X^m = (X_1, X_2, \ldots, X_m)$ be m i.i.d. repetitions from the distribution p_X.

Take some $\zeta > 0$. Let $k = \max_\nu e^{m(\zeta + D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}$ where the maximum is over all possible types ν (of sequences in $\mathcal{X}^m$). We now define $\hat{X}^m(1), \ldots, \hat{X}^m(k)$ jointly distributed with X^m. Given some x^m, we define the conditional distribution of $\hat{X}^m(1), \ldots, \hat{X}^m(k)$ given x^m as follows. Let $\nu(x)$ be the empirical type of the sequence x^m. For $j \leqslant e^{m(\zeta + D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}$, generate the sequences $\hat{X}^m(j)$ independently and i.i.d. from the distribution $q_\nu(\hat{x})$. For $j > e^{m(\zeta + D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}$, the sequences $\hat{X}^m(j)$ are all zero.

From Theorem 14 we get the following tail bound:

$$\mathbb{P}(X \geqslant \Delta)^m \leqslant \mathbb{P}\left(\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon\right) + \mathbb{P}\Big(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon\Big). \tag{72}$$

We have

$$\mathbb{P}\Big(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon\Big) = \sum_{x^m} p(x^m)\mathbb{P}\left(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon \middle| X^m = x^m\right).$$

Fix some $X^m = x^m$ with a type $\nu(x)$. Observe that if for some j, the sequence $\hat{X}^m(j)$ has type $p_\nu(\hat{x})$ then $\rho(X^m, \hat{X}^m(j)) \leqslant \epsilon$. This follows from the definition of p_ν in (70). Therefore, the probability $\mathbb{P}\left(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon \middle| X^m = x^m\right)$ is less than or equal to the probability that there is no j such that the sequence $\hat{X}^m(j)$ has type $p_\nu(\hat{x})$.

We now compute the probability that there is some $j \leqslant e^{m(\zeta + D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}$ such that the sequence $\hat{X}^m(j)$ has type $p_\nu(\hat{x})$. The probability that each sequence has type $p_\nu(\hat{x})$ is greater than or equal to (Csiszár, 1995, Lemma 2.6)

$$\alpha = (m+1)^{-|\hat{\mathcal{X}}|}e^{-mD_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x}))}.$$

The probability that there is no $j \leqslant e^{m(\zeta+D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}$ such that the sequence $\hat{X}^m(j)$ has type $p(\hat{x})$ equals

$$(1-\alpha)^{e^{m(\zeta+D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}} \leqslant \exp(-\alpha e^{m(\zeta+D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}) = \exp(-(m+1)^{-|\hat{\mathcal{X}}|}e^{m\zeta})$$

where we used the inequality $(1-x)^m \leqslant \exp(-mx)$. Since this upper bound does not depend on our choice of x^m, we get

$$\mathbb{P}\Big(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon\Big) \leqslant \exp(-(m+1)^{-|\hat{\mathcal{X}}|}e^{m\zeta}).$$

Next, note that

$$\mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon\right] = \sum_{x^m} p(x^m)\mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right].$$

For every $\epsilon < \Delta$, we can obtain an upper bound using the union bound as follows:

$$\begin{aligned}\mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right] &= \mathbb{P}\left[\exists j \leqslant e^{m(\zeta+D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))} : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right] \\ &\leqslant \sum_{j=1}^{e^{m(\zeta+D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}} \mathbb{P}\left[\frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right].\end{aligned}$$

Take some $j \leqslant e^{m(\zeta+D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x})))}$. Chernoff's bound implies that

$$\begin{aligned}\mathbb{P}\left[\frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right] &\leqslant \exp(-m\lambda(\Delta-\epsilon))\prod_{i=1}^m \mathbb{E}_{q_\nu}[\exp(\lambda \hat{X}_i(j))|x^m] \\ &= \exp(-m\lambda(\Delta-\epsilon) + m\log \mathbb{E}_{q_\nu}[\exp(\lambda \hat{X})]).\end{aligned}$$

We obtain

$$\begin{aligned}\mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right] &\leqslant \exp(m(\zeta + D_{KL}(p_\nu(\hat{x})\|q_\nu(\hat{x}))) - m\lambda(\Delta-\epsilon) + m\log \mathbb{E}_{q_\nu}[\exp(\lambda \hat{X})]) \\ &= \exp\Big[m\zeta - m\lambda(\Delta-\epsilon) + m\lambda \mathbb{E}_{p_\nu}(\hat{X})\Big]\end{aligned}$$

where we used (71) in the last step. Another trivial upper bound is

$$\mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right] \leqslant 1.$$

Thus,

$$\mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right] \leqslant \exp\left\{-m\Big[-\zeta + \lambda(\Delta-\epsilon) - \lambda \mathbb{E}_{p_\nu}(\hat{X})\Big]_+\right\}.$$

The above bound depends only on the type ν and not on the exact sequence x^m. If we denote $\mathcal{T}_\nu$ the set of sequences x^m with type ν, we have (Csiszár, 1995, Lemma 2.6)

$$\sum_{x^m \in \mathcal{T}_\nu} p(x^m) \leqslant \exp(-mD_{KL}(\nu_X \| p_X)).$$

Thus,

$$\begin{aligned}\sum_{x^m \in \mathcal{T}_\nu} p(x^m) \mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon \middle| x^m\right] \\ \leqslant \exp\left\{-mD_{KL}(\nu_X \| p_X) - m\left[-\zeta + \lambda(\Delta - \epsilon) - \lambda \mathbb{E}_{p_\nu}(\hat{X})\right]_+\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon\right] \\ &\leqslant \sum_\nu \exp\left\{-mD_{KL}(\nu_X \| p_X) - m\left[-\zeta + \lambda(\Delta - \epsilon) - \lambda \mathbb{E}_{p_\nu}(\hat{X})\right]_+\right\} \\ &\leqslant (m+1)^{|\mathcal{X}|} \max_\nu \exp\left\{-mD_{KL}(\nu_X \| p_X) - m\left[-\zeta + \lambda(\Delta - \epsilon) - \lambda \mathbb{E}_{p_\nu}(\hat{X})\right]_+\right\}.\end{aligned}$$

From (72) we get

$$\begin{aligned}\mathbb{P}(X \geqslant \Delta)^m &\leqslant \mathbb{P}\left[\exists j : \frac{1}{m}\sum_i \hat{X}_i(j) \geqslant \Delta - \epsilon\right] + \mathbb{P}\Big(\forall j \in [k] : \rho(X^m, \hat{X}^m(j)) > \epsilon\Big) \\ &\leqslant (m+1)^{|\mathcal{X}|} \max_\nu \exp\left\{-mD_{KL}(\nu_X \| p_X) - m\left[-\zeta + \lambda(\Delta - \epsilon) - \lambda \mathbb{E}_{p_\nu}(\hat{X})\right]_+\right\} \\ &\quad + \exp\Big(-(m+1)^{-|\hat{\mathcal{X}}|} e^{m\zeta}\Big).\end{aligned}$$

Raising both sides of the inequality to the power $1/m$ and letting m tend to infinity yields

$$\mathbb{P}(X \geqslant \Delta) \leqslant \max_\nu \exp\left\{-D_{KL}(\nu_X \| p_X) - \left[-\zeta + \lambda(\Delta - \epsilon) - \lambda \mathbb{E}_{p_\nu}(\hat{X})\right]_+\right\}.$$

Letting m tend to infinity, we obtain the above inequality for any arbitrary $p_\nu(\hat{x})$ in $\mathcal{P}(\nu_X)$. Letting ζ tend to zero yields the desired result. ■

E.16. Proof of Corollary 26

Proof

i. Let $\hat{W}$ be uniformly distributed over the d-dimensional ball with radius $\epsilon < R$ with center W. Let V_d denote the volume of the unitary d-dimensional ball. Then,

$$\begin{aligned} I(W;\hat{W}) =& H(\hat{W}) - H(\hat{W}|W) \\ =& H(\hat{W}) - \log(\epsilon^d V_d) \\ \leqslant& \log((r_0+\epsilon)^d V_d) - \log(\epsilon^d V_d) \\ =& d\log((r_0+\epsilon)/\epsilon) \\ \leqslant& d\log(2r_0/\epsilon). \end{aligned}$$

Now, using Corollary 11, we derive with probability at least $1-\delta$,

$$\operatorname{gen}(S,W) \leqslant \sqrt{\frac{2\sigma^2(d\log(2r_0/\epsilon)+\log(1/\delta))}{n}} + 2\mathfrak{L}\epsilon.$$

For $n \geqslant 16$, by letting $\epsilon \coloneqq 2r_0\sqrt{d\log(n)/n}$ and $\delta \coloneqq e^{-d/2}$, we derive that for $n \geqslant 16$, with probability at least $1-e^{-d/2}$,

$$\operatorname{gen}(S,W) \leqslant (4r_0\mathfrak{L} + \sigma\sqrt{d})\sqrt{\log(n)/n}.$$

ii. The result follows from (Marton, 1974, Example 1) and Corollary 11.

iii. The result follows from (Bakshi and Bansal, 2005, Example 1) and Corollary 11.

iv. The result follows from (Ihara and Kubo, 2000, Theorem 2) and Corollary 11.

■

E.17. Proof of Lemma 27

Proof Here we show the proof for $d_m \coloneqq \vartheta_m$. The proof is similar for $|\vartheta_m|$ and ξ_m. For simplicity, denote $\mathcal{E}_m \coloneqq \mathcal{E}_m(\mathcal{H}_m, \epsilon; \vartheta_m)$.

$$\begin{aligned} &\mathbb{E}_{(S,W)^{\otimes m}}\left(\min_{j\in[l_m]} \vartheta_m(W^m, \hat{\mathbf{w}}_j; S^m)\right) \\ &\quad= \mathbb{E}_{(S,W)^{\otimes m}}\left(\min_{j\in[l_m]} \vartheta_m(W^m, \hat{\mathbf{w}}_j; S^m)|\mathcal{E}_m\right)\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m) \\ &\qquad+ \mathbb{E}_{(S,W)^{\otimes m}}\left(\min_{j\in[l_m]} \vartheta_m(W^m, \hat{\mathbf{w}}_j; S^m)|\mathcal{E}_m^c\right)\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m^c) \\ &\quad\leqslant \mathbb{E}_{(S,W)^{\otimes m}}\left(\min_{j\in[l_m]} \vartheta_m(W^m, \hat{\mathbf{w}}_j; S^m)|\mathcal{E}_m\right)\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m) + \epsilon \\ &\quad\leqslant \mathbb{E}_{(S,W)^{\otimes m}}(\vartheta_m(W^m, \hat{\mathbf{w}}_1; S^m)|\mathcal{E}_m)\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m) + \epsilon \\ &\quad\leqslant \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left(|\sum_{i=1}^m \operatorname{gen}(S_i,W_i)| + |\sum_{i=1}^m \operatorname{gen}(S_i,\hat{w}_{1,i})|\Big|\mathcal{E}_m\right)\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m) + \epsilon \\ &\quad\overset{(a)}{\leqslant} \frac{1}{m}\mathbb{E}_{(S,W)^{\otimes m}}\left(|\sum_{i=1}^m \operatorname{gen}(S_i,W_i)|\Big|\mathcal{E}_m\right)\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{E}_m) + \sqrt{2\sigma^2\log(2)/(nm)} + \epsilon \\ &\quad\overset{(b)}{\leqslant} \varepsilon_m + \epsilon, \end{aligned}$$

where $\lim_{m\to\infty}\varepsilon_m = 0$, (a) is due to a known maximal inequality for subgaussian random variable (Boucheron et al., 2013, Theorem 2.5) and (b) is derived due to the Lemma 29, stated in the following. Taking the limit for $m \to \infty$ completes the proof.

To show step (b), we state the following lemma, shown within the proof of (Berger, 1975, Theorem 7.2.2). Here, for the sake of completeness, we state the adapted proof to our setup in Section E.19.

Lemma 29 *Assume that* $\mathbb{E}_X[|X|] < \infty$, *where* $X \in \mathcal{X}$. *Let* $\{\mathcal{A}_m\}_{m\in\mathbb{N}}\colon \mathcal{A}_m \subseteq \mathcal{X}^m$, *be a sequence of sets such that* $\lim_{m\to\infty}\mathbb{P}_{X^{\otimes m}}(X^m \in \mathcal{A}_m) = 0$. *Then,*

$$\lim_{m\to\infty}\frac{1}{m}\int_{\mathcal{A}_m}(\sum_{i=1}^{m}|X_i|)\,\mathrm{d}P_{X^{\otimes m}} = 0. \tag{73}$$

■

E.18. Proof of Lemma 28

Proof Denote that the marginal type of Q_m with respect to $\mathcal{S}^m$ as $Q_{s,m}$. Consider a random variable $\hat{W}$, defined by the conditional distribution $P_{\hat{W}|S}$, such that $\mathbb{E}\Big[d_{Q_m}(\hat{W};S)\Big] \leqslant \epsilon$, where the expectation is with respect to joint distribution $Q_{S,\hat{W}} \coloneqq Q_{s,m}\times P_{\hat{W}|S}$.

Following the proof of (Csiszár and Körner, 2011, Lemma 9.1), we can find a set $\mathcal{H}_{Q_m,\hat{W}} = \{\hat{\mathbf{w}}_{Q_m,j}, j\in[l_{Q_m,\hat{W}}]\}\in\hat{\mathcal{W}}^m$, such that

$$l_{Q_m,\hat{W}} \leqslant e^{m(I(S;\hat{W})+\varepsilon'_m)},$$

where the mutual information is with respect to the joint distribution $Q_{S,\hat{W}}$, and such that for each S^m having the type $Q_{s,m}$, there exists a $j(S^m)\in[l_{Q_m,\hat{W}}]$, such that

$$\|\hat{P}_{S^m,\hat{\mathbf{w}}_{Q_m,j(S^m)}}(s,\hat{w}) - Q_{S,\hat{W}}(s,\hat{w})\|_{TV} = \varepsilon''_m, \tag{74}$$

where ε''_m vanishes as $m\to\infty$. Note that by Carathéodory's theorem, we can assume that $\hat{\mathcal{W}}$ is a finite set as well and hence $d_{Q_m}(\hat{w};s)$ is always bounded. This yields for the picked $j(S^m)$, satisfying the above equation, we have

$$\sum_{i=1}^{m} d_{Q_m}(\hat{w}_{Q_m,j(S^m),i};S_i) \leqslant m(\epsilon+\varepsilon_m), \tag{75}$$

where ε_m vanishes as $m\to\infty$. Now,

$$\begin{aligned}\mathbb{P}_{(S,W)^{\otimes m}}&\left(\mathcal{T}(S^m,W^m) = Q_m, \min_{j\in[l_{Q_m,\hat{W}}]}\varphi_m(W^m,\hat{\mathbf{w}}_{Q_m,j};S^m) \geqslant \epsilon+\nu\right)\\ &=\mathbb{P}_{(S,W)^{\otimes m}}\left(\mathcal{T}(S^m,W^m) = Q_m, \min_{j\in[l_{Q_m,\hat{W}}]}\sum_{i=1}^{m}\min_{k\in[m]}\mathrm{gen}(S_k,W_k) - \mathrm{gen}(S_i,\hat{w}_{Q_m,j,i}) \geqslant m(\epsilon+\nu)\right)\end{aligned}$$

$$
\begin{aligned}
&=\mathbb{P}_{(S,W)^{\otimes m}}\left(\mathcal{T}(S^m,W^m)=Q_m,\min_{j\in[l_{Q_m,\hat{W}}]}\sum_{i=1}^{m}\min_{\substack{(s',w'):\\ Q_m(s',w')>0}}\operatorname{gen}(s',w')-\operatorname{gen}(S_i,\hat{w}_{Q_m,j,i})\geqslant m(\epsilon+\nu)\right)\\
&=\mathbb{P}_{(S,W)^{\otimes m}}\left(\mathcal{T}(S^m,W^m)=Q_m,\min_{j\in[l_{Q_m,\hat{W}}]}\sum_{i=1}^{m}d_{Q_m}(\hat{w}_{Q_m,j,i};S_i)\geqslant m(\epsilon+\nu)\right)\\
&\overset{(a)}{\leqslant}\mathbb{P}_{(S,W)^{\otimes m}}(\mathcal{T}(S^m,W^m)=Q_m,m(\epsilon+\varepsilon_m)\geqslant m(\epsilon+\nu))\\
&=0,
\end{aligned}
$$

where the last step holds for $m\geqslant m_{\nu,Q_m}$, where m_{ν,Q_m} is a sufficiently large integer.

The required set $\mathcal{H}_{Q_m}$ would be equal to the $\mathcal{H}_{Q_m,\hat{W}}$ having the minimum cardinality number $l_{Q_m,\hat{W}}$. This completes the proof. ■

E.19. Proof of Lemma 29

Proof Let $\eta\coloneqq\mathbb{E}_X[|X|]$ and $s_m(x^m)\coloneqq\frac{1}{m}\sum_{i=1}^{m}|x_i|$. Define the following sets

$$
\begin{aligned}
\mathcal{B}_m&\coloneqq\{x^m\colon s_m(x^m)>\eta+\delta\},\\
\mathcal{C}_m&\coloneqq\{x^m\colon s_m(x^m)<\eta-\delta\}.
\end{aligned}
$$

Fix a $\delta>0$. Then,

$$
\begin{aligned}
\int_{\mathcal{A}_m}S_m(X^m)\mathrm{d}P_{X^{\otimes m}}&=\int_{\mathcal{A}_m}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}+\eta\int_{\mathcal{A}_m}\mathrm{d}P_{X^{\otimes m}}\\
&=\int_{\mathcal{A}_m}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}+\eta\epsilon'_m\\
&=\int_{\mathcal{A}_m\bigcap\mathcal{B}_m}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}+\int_{\mathcal{A}_m\bigcap\mathcal{B}_m^c}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}+\eta\epsilon'_m\\
&=\int_{\mathcal{A}_m\bigcap\mathcal{B}_m}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}+\delta\int_{\mathcal{A}_m\bigcap\mathcal{B}_m^c}\mathrm{d}P_{X^{\otimes m}}+\eta\epsilon'_m\\
&\leqslant\int_{\mathcal{A}_m\bigcap\mathcal{B}_m}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}+\epsilon_m\\
&\overset{(a)}{\leqslant}\int_{\mathcal{B}_m}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}+\epsilon_m.
\end{aligned}
\tag{76}
$$

where ϵ_m vanishes as $m\to\infty$ and (a) is derived since for $X^m\in\mathcal{B}_m$, $S_m(X^m)-\eta$ is positive. Next,

$$
\begin{aligned}
\int_{\mathcal{B}_m}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}&=\int(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}-\int_{\mathcal{B}_m^c}(S_m(X^m)-\eta)\mathrm{d}P_{X^{\otimes m}}\\
&\overset{(a)}{=}\int_{\mathcal{B}_m^c}(\eta-S_m(X^m))\mathrm{d}P_{X^{\otimes m}}\\
&=\int_{\mathcal{B}_m^c\bigcap\mathcal{C}_m}(\eta-S_m(X^m))\mathrm{d}P_{X^{\otimes m}}+\int_{\mathcal{B}_m^c\bigcap\mathcal{C}_m^c}(\eta-S_m(X^m))\mathrm{d}P_{X^{\otimes m}}
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant \int_{\mathcal{B}_m^c \bigcap \mathcal{C}_m} (\eta - S_m(X^m)) \mathrm{d}P_{X^{\otimes m}} + \delta \\
&\overset{(b)}{\leqslant} \eta \int_{\mathcal{B}_m^c \bigcap \mathcal{C}_m} \mathrm{d}P_{X^{\otimes m}} + \delta \\
&= \eta \int_{\mathcal{C}_m} \mathrm{d}P_{X^{\otimes m}} + \delta \\
&\overset{(c)}{=} \epsilon_m'' + \delta,
\end{aligned} \tag{77}
$$

where ϵ_m'' vanishes as $m \to \infty$, (a) is derived since $\mathbb{E}[S_m(X^m)] = \eta$, (b) is derived since $S_m(X^m)$ is non-negative, and (c) is derived since $\mathbb{P}(|S_m - \eta| > \delta)$ asymptotically vanishes by law of large numbers.

The inequalities (76) and (77) yield for any $\delta > 0$,

$$
\lim_{m \to \infty} \frac{1}{m} \int_{\mathcal{A}_m} (\sum_{i=1}^{m} |X_i|) \, \mathrm{d}P_{X^{\otimes m}} < \delta. \tag{78}
$$

This completes the proof. ■