Tighter Expected Generalization Error Bounds via Wasserstein Distance

Borja Rodríguez-Gálvez
KTH Royal Institute of Technology
Stockholm, Sweden
borjarg@kth.se

Germán Bassi
Ericsson Research
Stockholm, Sweden
german.bassi@ericsson.com

Ragnar Thobaben
KTH Royal Institute of Technology
Stockholm, Sweden
ragnart@kth.se

Mikael Skoglund
KTH Royal Institute of Technology
Stockholm, Sweden
skoglund@kth.se

Abstract

This work presents several expected generalization error bounds based on the Wasserstein distance. More specifically, it introduces full-dataset, single-letter, and random-subset bounds, and their analogues in the randomized subsample setting from Steinke and Zakynthinou [1]. Moreover, when the loss function is bounded and the geometry of the space is ignored by the choice of the metric in the Wasserstein distance, these bounds recover from below (and thus, are tighter than) current bounds based on the relative entropy. In particular, they generate new, non-vacuous bounds based on the relative entropy. Therefore, these results can be seen as a bridge between works that account for the geometry of the hypothesis space and those based on the relative entropy, which is agnostic to such geometry. Furthermore, it is shown how to produce various new bounds based on different information measures (e.g., the lautum information or several f-divergences) based on these bounds and how to derive similar bounds with respect to the backward channel using the presented proof techniques.

1 Introduction

A *learning algorithm* is a mechanism that takes a *dataset* $s = (z_1, \dots, z_n)$ of n samples $z_i \in \mathcal{Z}$ taken i.i.d. from a distribution P_Z as an input, and produces a hypothesis $w \in \mathcal{W}$ by means of the conditional probability distribution $P_{W|S}$.

The ability of a hypothesis w to characterize a sample z is described by the loss function $\ell(w, z) \in \mathbb{R}$. More precisely, a hypothesis w describes well the samples from a population P_Z when its *population risk*, i.e., $\mathscr{L}_{P_Z}(w) \triangleq \mathbb{E}[\ell(w, Z)]$, is low. However, the distribution P_Z is often not available and the *empirical risk* on the dataset s, i.e., $\mathscr{L}_s(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$, is considered as a proxy. Therefore, it is of interest to study the discrepancy between the population and empirical risks, which is defined as the *generalization error*:

$$\operatorname{gen}(w, s) \triangleq \mathscr{L}_{P_Z}(w) - \mathscr{L}_s(w).$$

Classical approaches bound the generalization error in expectation and in probability (PAC Bayes) either by studying the complexity and the geometry of the hypothesis' space $\mathcal{W}$ or by exploring properties of the learning algorithm itself; see, e.g., [2, 3] for an overview.

More recently, the relationship (or amount of information) between the generated hypothesis and the training dataset has been used as an indicator of the generalization performance. In [4], based on

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

[5], it is shown that the expected generalization error, i.e., $\overline{\mathrm{gen}}(W,S) \triangleq \mathbb{E}[\mathrm{gen}(W,S)]$, is bounded from above by a function that depends on the mutual information between the hypothesis W and the dataset S with which it is trained, i.e., $I(W;S)$. However, this bound becomes vacuous when $I(W;S) \to \infty$, which occurs for example when W and S are separately continuous and W is a deterministic function of S. To address this issue, it is shown in [6] that the generalization error is also bounded by a function on the dependency between the hypothesis and individual samples, $I(W;Z_i)$, which is usually finite due to the smoothing effect of marginalization. Following this line of work, in [7], the authors present data-dependent bounds based on the relationship between the hypothesis and random subsets of the data, i.e., $D_{\mathrm{KL}}(P_{W|s} \,\|\, P_{W|s_{j^c}})$ where $j \subseteq [n]$.

After that, a more structured setting is introduced in [1], studying instead the relationship between the hypothesis and the *identity* of the samples. The authors consider a super-sample of $2n$ i.i.d. instances $\tilde{z}_i$ from P_Z, i.e., $\tilde{s} = (\tilde{z}_1, \ldots, \tilde{z}_{2n})$. This super sample is used to construct the dataset s by choosing between the samples $\tilde{z}_i$ and $\tilde{z}_{i+n}$ using a Bernoulli random variable U_i with probability $\frac{1}{2}$, i.e., $z_i = \tilde{z}_{i+u_i n}$. In this paper, the two settings are referred to as the *standard* and *randomized-subsample* settings.[1] In the randomized-subsample setting, the *empirical generalization error* is defined as the difference between the empirical risk on the samples from $\tilde{s}$ not used to obtain the hypothesis, i.e., $\bar{s} = \tilde{s} \setminus s$, and the empirical risk on the dataset s, i.e.,

$$\widehat{\mathrm{gen}}(w,\tilde{s},u) \triangleq \mathscr{L}_{\bar{s}}(w) - \mathscr{L}_s(w) = \frac{1}{n}\sum\nolimits_{i=1}^{n} \big(\ell(w,\tilde{z}_{i+(1-u_i)n}) - \ell(w,\tilde{z}_{i+u_i n})\big),$$

where u is the sequence of n i.i.d. Bernoulli trial outcomes u_i. The expected value of the empirical and the (standard) generalization errors coincide, i.e., $\mathbb{E}[\widehat{\mathrm{gen}}(W,\tilde{S},U)] = \overline{\mathrm{gen}}(W,S)$. Also, the expected generalization error is controlled by the conditional mutual information between the hypothesis W and the Bernoulli trials U, given the super sample $\tilde{S}$ [1], i.e., $I(W;U|\tilde{S})$, by the individual conditional mutual information [9], i.e., $I(W;U_i|\tilde{Z}_i,\tilde{Z}_{i+n})$, and by the "disintegrated" mutual information with a subset U_J of the Bernoulli trials [10], i.e., $D_{\mathrm{KL}}(P_{W|\tilde{S},U} \,\|\, P_{W|\tilde{S},U_{J^c}})$. A highlight of this setting is that these conditional notions of information are always finite [1] and smaller than their "unconditional" counterparts [10], e.g., $I(W;U|\tilde{S}) \leq I(W;S)$ and $I(W;U|\tilde{S}) \leq n\log(2)$.

Some steps towards unifying these results are taken in [8], where the authors develop a framework that makes it possible to recover the expected generalization error bounds based on the mutual information $I(W;S)$ and the conditional mutual information $I(W;U|\tilde{S})$. Then, the aforementioned framework is further exploited in [9] to recover the single-sample and the random-subsets bounds, which are based on $I(W;Z_i)$, $D_{\mathrm{KL}}(P_{W|S} \,\|\, P_{W|S_{J^c}})$, and $D_{\mathrm{KL}}(P_{W|\tilde{S},U} \,\|\, P_{W|\tilde{S},U_{J^c}})$, and to generate new individual conditional mutual information bounds, i.e., $I(W;U_i|\tilde{Z}_i,\tilde{Z}_{i+n})$. Finally, in [11, 12], other systematic ways to recover some of the said bounds and obtain similar new ones are studied.

In parallel, there were some attempts to bridge the gap between employing the geometry and complexity of the hypothesis space and the relationship between the hypothesis and the training samples. In [13], the authors bound $\overline{\mathrm{gen}}(W,S)$ with a function of weighted dependencies between the dataset and increasingly finer quantizations of the hypothesis, i.e., $\{2^{-k/2} I([W]_k;S)\}_k$, which can be finite even if $I(W;S) \to \infty$. This result stems from a clever usage of the chaining technique [14, Theorem 5.24], and a comparison with this kind of approaches is given in Appendix A. Later, in [15] and [16], it is shown that the expected generalization error is bounded from above by a function of the Wasserstein distance between the hypothesis distribution after observing the dataset $P_{W|S}$ and its prior P_W, i.e., $\mathbb{W}_p(P_{W|S}, P_W)$, and by a function of the Wasserstein distance between the hypothesis distribution after observing a single sample $P_{W|Z_i}$ and its prior P_W, i.e., $\mathbb{W}_p(P_{W|Z_i}, P_W)$, which are finite when a suitable metric is chosen but are difficult to evaluate. Concurrently, in [17] it is shown that a similar result holds, if the metric is the Minkowski distance, for the distribution of the data P_S and the backward channel $P_{S|W}$, i.e., $\mathbb{W}^p_{p,\|\cdot\|}(P_{S|W}, P_S)$.

The main contributions of this paper are the following:

- It introduces new, tighter single letter and random-subset Wasserstein distance bounds for the standard and randomized-subsample settings (Theorems 1, 2, 3, and 4).
- It shows that when the loss is bounded and the geometry of the space is ignored, these bounds recover from below (and thus are tighter than) the current relative entropy and mutual

[1] In [8] the latter is called the random-subset setting. However, this may cause confusion with the random-subset bounds in the present work.

information bounds on both the standard and randomized-subsample settings. In fact, they are also tighter when the loss is additionally subgaussian or under certain milder conditions on the geometry. However, these results are deferred to Appendix B to expose the main ideas more clearly. Moreover, Corollaries 1 and 2 overcome the issue of potentially vacuous relative entropy bounds on the standard setting.

- It introduces new bounds based on the backward channel, which are analogous to those based on the forward channel and more general than previous results in [17].
- It shows how to generate new bounds based on a variety of information measures, e.g., the lautum information or several f-divergences like the Hellinger distance or the χ^2-divergence, thus making the characterization of the generalization more flexible.

2 Preliminaries

2.1 Notation

Random variables X are written in capital letters, their realizations x in lower-case letters, their set of outcomes $\mathcal{X}$ in calligraphic letters, and their Borel σ-algebras $\mathfrak{X}$ in script-style letters. Moreover, the probability distribution of a random variable X is written as $P_X : \mathfrak{X} \to [0,1]$. Hence, the random variable X or the probability distribution P_X induce the probability space $(\mathcal{X}, \mathfrak{X}, P_X)$. When more than one random variable is considered, e.g., X and Y, their joint distribution is written as $P_{X,Y} : \mathfrak{X} \otimes \mathcal{Y} \to [0,1]$ and their product distribution as $P_X \otimes P_Y : \mathfrak{X} \otimes \mathcal{Y} \to [0,1]$. Moreover, the conditional probability distribution of Y given X is written as $P_{Y|X} : \mathcal{Y} \otimes \mathcal{X} \to [0,1]$ and defines a probability distribution $P_{Y|X=x}$ (or $P_{Y|x}$ for brevity) over $\mathcal{Y}$ for each element $x \in \mathcal{X}$. Finally, there is an abuse of notation writing $P_{X,Y} = P_{Y|X} \times P_X$ since $P_{X,Y}(B) = \int \big(\int \chi_B\big((x,y)\big) dP_{Y|X=x}(y)\big) dP_X(x)$ for all $B \in \mathfrak{X} \otimes \mathcal{Y}$, where χ_B is the characteristic function of the set B. The natural logarithm is $\log$.

2.2 Necessary definitions, remarks, claims, and lemmas

Definition 1. *Let $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a metric. A space $(\mathcal{X}, \rho)$ is Polish if it is complete and separable. Throughout it is assumed that all Polish spaces $(\mathcal{X}, \rho)$ are equipped with the Borel σ-algebra $\mathfrak{X}$ generated by ρ. When there is no ambiguity, both the metric space $(\mathcal{X}, \rho)$ and the generated measurable space $(\mathcal{X}, \mathfrak{X})$ are written as $\mathcal{X}$.*

Definition 2. *Let $(\mathcal{X}, \rho)$ be a Polish metric space and let $p \in [1, \infty)$. Then, the* Wasserstein distance *of order p between two probability distributions P and Q on $\mathcal{X}$ is*

$$\mathbb{W}_p(P,Q) \triangleq \Big(\inf_{R \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{X}} \rho(x,y)^p dR(x,y) \Big)^{1/p},$$

where $\Pi(P,Q)$ is the set of all couplings R of P and Q, i.e., all joint distributions on $\mathcal{X} \times \mathcal{X}$ with marginals P and Q, that is, $P(B) = R(B, \mathcal{X})$ and $Q(B) = R(\mathcal{X}, B)$ for all $B \in \mathfrak{X}$.

Remark 1. *Hölder's inequality implies that $\mathbb{W}_p \leq \mathbb{W}_q$ for all $p \leq q$ [18, Remark 6.6]. Hence, since this work is centered on upper bounds the focus is on $\mathbb{W} \triangleq \mathbb{W}_1$.*

Definition 3. *A function $f : \mathcal{X} \to \mathbb{R}$ is said to be L-Lipschitz under the metric ρ, or simply $f \in L\text{-Lip}(\rho)$, if $|f(x) - f(y)| \leq L\rho(x,y)$ for all $x, y \in \mathcal{X}$.*

Lemma 1 (Kantorovich-Rubinstein duality [18, Remark 6.5])**.** *Let $\mathcal{P}_1(\mathcal{X})$ be the space of probability distributions on $\mathcal{X}$ with a finite first moment. Then, for any two distributions P and Q in $\mathcal{P}_1(\mathcal{X})$*

$$\mathbb{W}(P,Q) = \sup_{f \in 1\text{-Lip}(\rho)} \Big\{ \int_{\mathcal{X}} f(x) dP(x) - \int_{\mathcal{X}} f(x) dQ(x) \Big\}. \qquad \text{(KR duality)}$$

Definition 4. *The total variation between two probability distributions P and Q on $\mathcal{X}$ is*

$$\mathtt{TV}(P,Q) \triangleq \sup_{A \in \mathfrak{X}} \{P(A) - Q(A)\}.$$

Definition 5. *The discrete metric is $\rho_{\mathrm{H}}(x,y) \triangleq \mathbb{1}[x \neq y]$, where $\mathbb{1}$ is the indicator function.*

Remark 2. *A bounded function $f : \mathcal{X} \to [a,b]$ is $(b-a)$-Lipschitz under the discrete metric ρ_{H}.*

Remark 3. *The Wasserstein distance of order 1 is dominated by the total variation. For instance, if P and Q are two distributions on $\mathcal{X}$ then $\mathbb{W}(P,Q) \leq d_\rho(\mathcal{X})\mathtt{TV}(P,Q)$, where $d_\rho(\mathcal{X})$ is the diameter of $\mathcal{X}$. In particular, when the discrete metric is considered $\mathbb{W}(P,Q) = \mathtt{TV}(P,Q)$ [18, Theorem 6.15].*

Lemma 2 (Pinsker's and Bretagnolle–Huber's (BH) inequalities)**.** *Let P and Q be two probability distributions on $\mathcal{X}$ and define $\Psi(x) \triangleq \sqrt{\min\{x/2, 1-\exp(-x)\}}$, then [19, Theorem 6.5] and [20, Proof of Lemma 2.1] state that*

$$\mathtt{TV}(P,Q) \leq \Psi(D_{\mathrm{KL}}(P \,\|\, Q)).$$

3 Expected Generalization Error Bounds

This section presents our main results. First, in §3.1 and §3.2, single-letter and random-subset bounds based on the Wasserstein distance are introduced for the studied settings. These subsections also show how these bounds are tighter than current bounds based on the Wasserstein distance and the relative entropy. Moreover, an example where these bounds outperform current bounds is provided. Then, in §3.3 it is shown how to obtain analogous bounds to those in §3.1 and §3.2 for the backward channel. Finally, §3.4 shows how the presented results lead to a rich set of new bounds based on different information measures. All complete proofs and technical details are deferred to the appendix.

3.1 Standard setting

In [15, Theorem 2], the authors show that the expected generalization error is bounded from above by the Wasserstein distance between the forward channel distribution $P_{W|S}$ and the marginal distribution of the hypothesis P_W. More specifically, when the loss function ℓ is L-Lipschitz under a metric ρ for all $z \in \mathcal{Z}$ and the hypothesis space $\mathcal{W}$ is Polish, then

$$\left|\overline{\mathrm{gen}}(W,S)\right| \leq L\mathbb{E}\left[\mathbb{W}(P_{W|S}, P_W)\right] = L\int_{\mathcal{Z}^n} \mathbb{W}(P_{W|S=s}, P_W) dP_Z^{\otimes n}(s). \tag{1}$$

This bound considers both the geometry of the hypothesis space by means of the metric ρ and the dependence between the hypothesis and the dataset via the discrepancy between the forward channel $P_{W|S}$ and the marginal P_W. Nonetheless, it is not clear how it relates with other results agnostic to the geometry of the space. For instance, when the loss function ℓ is bounded in $[a,b]$, if the geometry is ignored (i.e., the discrete metric is considered), then

$$\left|\overline{\mathrm{gen}}(W,S)\right| \leq (b-a)\mathbb{E}\left[\mathtt{TV}(P_{W|S}, P_W)\right] \leq (b-a)\Psi(I(W;S)),$$

where the inequalities follow from Remark 2, Lemma 2, and Jensen's inequality (note $\Psi(x)$, defined in Lemma 2, is concave on x). This result compares negatively with other results employing the mutual information, e.g., [4, Theorem 1], where the bound has a decaying factor of $1/\sqrt{n}$.

Nonetheless, it is possible to find a single-letter version of [15, Theorem 2] using a similar strategy to [6, Proposition 1] and [9, Propositions 1 and 3], which generalizes [16, Theorem 1] to algorithms that may consider the ordering of the samples. More concretely, the expected generalization error is controlled by a function of the Wasserstein distance of the hypothesis' distribution before and after observing a *single sample* Z_i, i.e., $\mathbb{W}(P_{W|Z_i}, P_W)$.

Theorem 1. *Suppose that the loss function ℓ is L-Lipschitz for all $z \in \mathcal{Z}$ and that the hypothesis space $\mathcal{W}$ is Polish. Then,*

$$\left|\overline{\mathrm{gen}}(W,S)\right| \leq \frac{L}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{W}(P_{W|Z_i}, P_W)\right].$$

Moreover, when the loss function is bounded and the geometry of the space is ignored by considering the discrete metric, this single-letter result can improve upon current relative entropy and mutual information bounds.

Corollary 1. *Under the conditions of Theorem 1, if the loss ℓ is bounded in $[a,b]$, then*

$$\left|\overline{\mathrm{gen}}(W,S)\right| \leq \frac{b-a}{n}\sum\nolimits_{i=1}^{n} \mathbb{E}\left[\mathtt{TV}(P_{W|Z_i}, P_W)\right] \leq \frac{b-a}{n}\sum\nolimits_{i=1}^{n} \mathbb{E}\left[\Psi\left(D_{\mathrm{KL}}(P_{W|Z_i} \,\|\, P_W)\right)\right]$$

Corollary 1 improves upon [6, Proposition 1] in two different ways. First, it pulls the expectation with respect to the samples P_{Z_i} outside of the concave square root, thus strengthening that result via Jensen's inequality. Second, the addition of the BH inequality ensures that heavily influential samples (high $I(W;Z_i)$) do not contribute too negatively to the bound, which is ensured to be non-vacuous. Moreover, contrarily to (1), a further application of Jensen's inequality and [6, Proposition 2] indicates that Corollary 1 compares positively to [4, Theorem 1], exhibiting the decaying factor of $1/\sqrt{n}$,

$$\left|\overline{\mathrm{gen}}(W,S)\right| \leq \frac{b-a}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{W}(P_{W|Z_i}, P_W)\right] \leq (b-a)\Psi\left(\frac{I(W;S)}{n}\right) \leq \sqrt{\frac{(b-a)^2 I(W;S)}{2n}}. \tag{2}$$

It is also possible to obtain a random-subset version of [15, Theorem 2] using a similar strategy to [9, Propositions 2 and 4]. This kind of bounds, rather than looking at how knowing a *single sample* Z_i modifies the hypothesis distribution, i.e., $\mathbb{W}(P_{W|Z_i}, P_W)$, look at how the knowledge of a set of samples S_J alters the hypothesis distribution when all the other samples, S_{J^c}, used to obtain the hypothesis are known too, i.e., $W(P_{W|S}, P_{W|S_{J^c}})$.

Theorem 2. *Suppose that the loss function ℓ is L-Lipschitz for all $z \in \mathcal{Z}$ and that the hypothesis space $\mathcal{W}$ is Polish. Let J be a uniformly random subset of $[n]$ such that $|J| = m$, and that is independent of W and S. Let also R be a random variable independent of S and J. Then,*

$$\left|\overline{\operatorname{gen}}(W,S)\right| \leq L\mathbb{E}\left[\mathbb{W}(P_{W|S,R}, P_{W|S_{J^c},R})\right] \text{ and}$$
$$\left|\overline{\operatorname{gen}}(W,S)\right| \leq \frac{L}{m}\mathbb{E}\left[\sum_{i\in J}\mathbb{E}\left[\mathbb{W}(P_{W|S_{J^c}\cup Z_i,R}, P_{W|S_{J^c},R}) \mid J\right]\right].$$

In particular, when $m = 1$, the two equations from Theorem 2 reduce to [21, Lemma 3]

$$\left|\overline{\operatorname{gen}}(W,S)\right| \leq L\mathbb{E}\left[\mathbb{W}(P_{W|S,R}, P_{W|S^{-J},R})\right],$$

where $S^{-J} = S \setminus Z_J$, i.e., the whole dataset except sample Z_J. Moreover, if the loss is bounded and the geometry is ignored, Theorem 2 improves upon the tightest bounds in terms of the relative entropy of random subsets, cf. [7, Theorem 2.5].

Corollary 2. *In the conditions of Theorem 2, if the loss is bounded in $[a, b]$, then*

$$\left|\overline{\operatorname{gen}}(W,S)\right| \leq (b-a)\mathbb{E}\left[\mathtt{TV}(P_{W|S,R}, P_{W|S^{-J},R})\right] \leq \frac{b-a}{n}\sum_{i=1}^{n}\mathbb{E}\left[\Psi\left(D_{\mathrm{KL}}(P_{W|S,R} \,\|\, P_{W|S^{-J},R})\right)\right].$$

These data-dependent bounds characterize well the expected generalization error of the Langevin dynamics (LD) and stochastic gradient Langevin dynamics (SGLD) algorithms [7, Theorems 3.1 and 3.3], where R is an artificial random variable used to encode some knowledge necessary to characterize the hypothesis distribution, such as the batch indices of SGLD. In particular, Corollary 2 improves upon [7, Theorem 2.5] tightening the elements of the expectation with respect to J for which the divergence is large ($\gtrapprox 1.6$).

It is possible to prove that Theorem 1 is tighter than [15, Theorem 1]. This results by studying the KR dual representation of the Wasserstein distance and noting that the conditional distribution $P_{W|Z_i}$ is a smoothed version of the forward channel, i.e., $P_{W|Z_i} = \mathbb{E}[P_{W|S}|Z_i]$. Comparisons with Theorem 2 are also possible using similar arguments and the triangle inequality. These results are informally summarized below and presented with more details and the proofs in Appendix D.1.

Proposition 1. *Consider the standard setting. Then, for all $j \subseteq [n]$ and all $i \in j$:*

$$\mathbb{E}\left[\mathbb{W}(P_{W|Z_i}, P_W)\right] \leq \mathbb{E}\left[\mathbb{W}(P_{W|S}, P_W)\right], \text{where } j = [n], \quad (\implies \text{Theorem 1} \leq [15, \text{Theorem 1}])$$
$$\mathbb{E}\left[\mathbb{W}(P_{W|Z_i}, P_W)\right] \leq \mathbb{E}\left[\mathbb{W}(P_{W|S}, P_{W|S_{j^c}})\right], \text{and} \quad (\implies \text{Theorem 1} \leq \text{Theorem 2})$$
$$\mathbb{E}\left[\mathbb{W}(P_{W|S}, P_{W|S_{j^c}})\right] \leq 2\mathbb{E}\left[\mathbb{W}(P_{W|S}, P_W)\right]. \quad (\implies \text{Theorem 2} \leq 2\cdot[15, \text{Theorem 1}])$$

The following example showcases a situation where the presented bounds outperform the current known bounds based on the Wasserstein distance and the mutual information.

Example 1 (Gaussian location model)**.** *Consider the problem of estimating the mean μ of a d-dimensional Gaussian distribution with known covariance matrix $\sigma^2 I_d$. Further consider that there are n samples $S = (Z_1, \ldots, Z_n)$ available, the loss is measured with the Euclidean distance $\ell(w, z) = \|w - z\|_2$, and the estimation is their empirical mean $W = \frac{1}{n}\sum_{i=1}^{n} Z_i$.*

In this example, the expected generalization error can be calculated exactly (see Appendix E):

$$\overline{\operatorname{gen}}(W,S) = \sqrt{\frac{2\sigma^2}{n}}\left(\sqrt{n+1} - \sqrt{n-1}\right)\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \in \mathcal{O}\left(\frac{\sqrt{\sigma^2 d}}{n}\right).$$

As discussed in [6], the bound from [4] is not applicable in this setting since $I(W;S) \to \infty$ and since $\ell(w, Z)$ is not subgaussian given that $\operatorname{Var}[\ell(w, Z)] \to \infty$ as $\|w\|_2 \to \infty$. When $d = 1$, the loss $\ell(W, Z)$ is 1-subgaussian and the individual sample mutual information (ISMI) bound from [6] produces a bound in $\mathcal{O}\left(\sqrt{\sigma^2/n}\right)$, which decreases slower than the true generalization error, see Figure 1. This happens since the bound grows as the square root of $I(W; Z_i)$, which is in $\mathcal{O}(1/n)$.

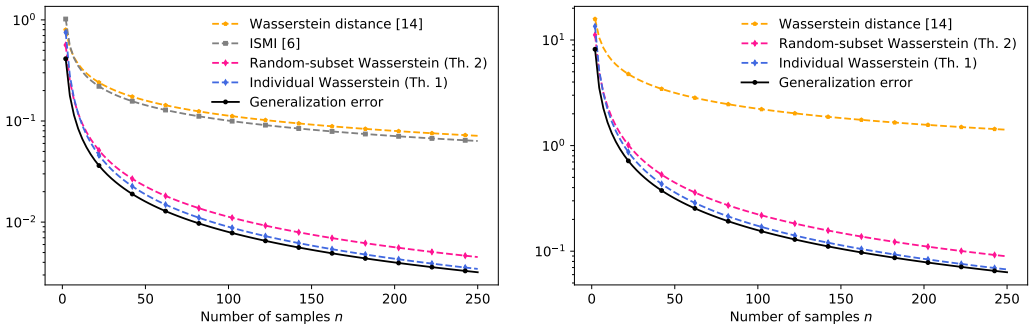

Figure 1: Expected generalization error and generalization error bounds for the Gaussian location model with $\mathcal{N}(\mu, 1)$ (left) and $\mathcal{N}(\mu, I_{250})$ (right). See Appendix E for the details.

In this scenario, the loss is 1-Lipschitz under $\rho(w, w') = \|w - w'\|_2$, and thus the bounds based on the Wasserstein distance are applicable. Applying the bound from [15] yields a bound in $\mathcal{O}\big(\sqrt{\sigma^2 d/n}\big)$, which decreases at the same sub-optimal rate as the ISMI bound. However, both the individual and random-subset Wasserstein distance bounds from Theorems 1 and 2 produce bounds in $\mathcal{O}\big(\sqrt{\sigma^2 d}/n\big)$, which decrease at the same rate as the true generalization error (see Figure 1).

3.1.1 Outline of the proofs

Similarly to [15, 16], the proofs of the theorems in this section are based on operating with $\overline{\mathrm{gen}}(W, S)$ until an expression of the type $\mathbb{E}[f(X', Y) - f(X, Y)]$ is reached, where X' is an independent copy of X such that $P_{X',Y} = P_X \otimes P_Y$, and then applying the KR duality. For example, in Theorem 1 such an expression is achieved with $X = W$, $Y = Z_i$, and $f = \ell$. To arrive at these expressions, the proofs of Theorem 3 and 4 operate with $\overline{\mathrm{gen}}(W, S)$ in different forms. More precisely,

(Th. 1) Since the samples Z_i are independent and the expectation is a linear operator, the proof follows working with the quantity $\overline{\mathrm{gen}}(W, S) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(W', Z_i) - \ell(W, Z_i)]$.

(Th. 2) Note that $\mathbb{E}[\mathscr{L}_{s_J}(w)] = \mathscr{L}_s(w)$, where J is a uniformly random subset of $[n]$ of size m and s_J is the subset of s indexed by J. This equality follows since there are $\binom{n}{m}$ subsets of size m and each sample z_i belongs to only $\binom{n-1}{m-1}$ of them. Hence,

$$\mathbb{E}[\mathscr{L}_{s_J}(w)] = \frac{1}{\binom{n}{m}} \sum\nolimits_{j \in \mathcal{J}} \frac{1}{m} \sum_{i \in j} \ell(w, z_i) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i) = \mathscr{L}_s(w).$$

Then, the proof follows working with the quantity $\overline{\mathrm{gen}}(W, \tilde{S}) = \mathbb{E}[\mathscr{L}_{S_J}(W') - \mathscr{L}_{S_J}(W)]$.

3.2 Randomized-subsample setting

In the randomized-subsample setting the focus shifts from studying the impact of the samples on the hypothesis distribution to the impact of the samples' identities on the hypothesis distribution. For example, the analogous result to (1) is

$$\big|\overline{\mathrm{gen}}(W, S)\big| \leq 2L\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})\big]. \tag{3}$$

Similarly to the standard setting, considering the discrete metric and applying Pinsker's and Jensen's inequalities leads to a less favorable bound than current bounds based on the mutual information [1, Theorem 5.1] since it does not explicitly decrease as $1/\sqrt{n}$. However, the bound still admits a tighter (see Appendix D.1) single-letter version.

Theorem 3. *Suppose that the loss function ℓ is L-Lipschitz for all $z \in \mathcal{Z}$ and that the hypothesis space $\mathcal{W}$ is Polish and let $\tilde{S}_i \triangleq (\tilde{Z}_i, \tilde{Z}_{i+n})$. Then,*

$$\big|\overline{\mathrm{gen}}(W, S)\big| \leq \frac{2L}{n} \sum_{i=1}^{n} \mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}_i})\big].$$

As in the standard setting, when the loss function is bounded and the geometry of the space is ignored, this result improves upon current single-letter bounds based on the mutual information [9, 12] by

pulling the expectation with respect to the samples $P_{\tilde{S}_i}$ out of the square root. Here, the BH inequality is not considered since $D_{\mathrm{KL}}(P_{W|\tilde{S}_i,U_i} \,\|\, P_{W|\tilde{S}_i}) \leq \log(2)$; see Appendix F for the details.

Corollary 3. *Under the conditions of Theorem 3, if the loss ℓ is bounded in $[a,b]$, then*

$$\begin{aligned}\left|\overline{\mathrm{gen}}(W,S)\right| &\leq \frac{2(b-a)}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathtt{TV}(P_{W|\tilde{S}_i,U_i},P_{W|\tilde{S}_i})\right]\\ &\leq \frac{b-a}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sqrt{2D_{\mathrm{KL}}(P_{W|\tilde{S}_i,U_i}\,\|\,P_{W|\tilde{S}_i})}\right].\end{aligned}$$

In this setting, Corollary 3 also decreases at a $1/\sqrt{n}$ rate and is tighter than [1, Theorem 5.1].

$$\left|\overline{\mathrm{gen}}(W,S)\right| \leq \frac{2(b-a)}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{W}(P_{W|\tilde{Z}_i,\tilde{Z}_{i+n},U_i},P_{W|\tilde{Z}_i,\tilde{Z}_{i+n}})\right] \leq \sqrt{\frac{2(b-a)^2 I(W;U|\tilde{S})}{n}}.$$

Finally, the randomized-subsample setting also accepts random-subset bounds. These bounds study how the knowledge of the *identities of a set of samples* that where used for training, U_J, alters the hypothesis distribution when all the other identities, U_{J^c}, and all the samples, $\tilde{S}$, are known.

Theorem 4. *Suppose that the loss function ℓ is L-Lipschitz for all $z \in \mathcal{Z}$ and that the hypothesis space $\mathcal{W}$ is Polish. Let J be a uniformly random subset of $[n]$ such that $|J| = m$, and that is independent of W, $\tilde{S}$, and ULet also R be a random variable independent of $\tilde{S}$,U, and J. Then,*

$$\begin{aligned}\left|\overline{\mathrm{gen}}(W,S)\right| &\leq 2L\mathbb{E}\left[\mathbb{W}(P_{W|\tilde{S},U,R},P_{W|\tilde{S},U_{J^c},R})\right] \text{ and}\\ \left|\overline{\mathrm{gen}}(W,S)\right| &\leq \frac{2L}{m}\mathbb{E}\Big[\sum_{i\in J}\mathbb{E}\left[\mathbb{W}(P_{W|\tilde{S},U_{J^c}\cup U_i,R},P_{W|\tilde{S},U_{J^c},R})\mid J\right]\Big].\end{aligned}$$

Although these bounds are weaker than Theorem 3 (see Appendix D.1), their data-dependent nature may lead to more tractable and sharper bounds in practice. For example, when the discrete metric is considered, Theorem 4 recovers from below current random-subset bounds based on the relative entropy, which are used to obtain some of the tightest bounds for LD and SGLD [9, 10].

Corollary 4. *In the conditions of Theorem 4, for $m = 1$, if the loss is bounded in $[a,b]$, then*

$$\begin{aligned}\left|\overline{\mathrm{gen}}(W,S)\right| &\leq 2(b-a)\mathbb{E}\left[\mathtt{TV}(P_{W|\tilde{S},U,R},P_{W|\tilde{S},U^{-J},R})\right]\\ &\leq \frac{b-a}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sqrt{2D_{\mathrm{KL}}(P_{W|\tilde{S},U,R}\,\|\,P_{W|\tilde{S},U^{-J},R})}\right].\end{aligned}$$

3.2.1 Outline of the proofs

The proofs of the results in this section are similar to those of the standard setting, hence their similar expressions. However, instead of operating with the expected generalization error in the form of $\mathbb{E}[\mathrm{gen}(W,S)]$ they operate with $\mathbb{E}[\widehat{\mathrm{gen}}(W,\tilde{S},U)]$.

There are two issues that complicate the application of the KR duality as in the previous proofs. For instance, consider $\widehat{\mathrm{gen}}(W,\tilde{S},U)$, then:

- Both $\mathscr{L}_{\bar{S}}(W)$ and $\mathscr{L}_S(W)$ depend on P_U. Hence, considering a copy W' of W such that $P_{W',\tilde{S},U} = P_{W,\tilde{S}} \otimes P_U$ does not help since $\mathbb{E}[\widehat{\mathrm{gen}}(W,\tilde{S},U)] \neq \mathbb{E}[\mathscr{L}_{\bar{S}}(W') - \mathscr{L}_S(W)]$.
- Even if $\mathbb{E}[\widehat{\mathrm{gen}}(W,\tilde{S},U)] = \mathbb{E}[\mathscr{L}_{\bar{S}}(W') - \mathscr{L}_S(W)]$ were true, for some fixed $\tilde{s}$ and u, the functions $\mathscr{L}_{\tilde{s}}(w)$ and $\mathscr{L}_s(w)$ on w are different, and thus the KR duality cannot be invoked.

Nonetheless, these two issues are resolved considering instead

$$\widehat{\mathrm{gen}}(W,\tilde{S},U) = \mathscr{L}_{\bar{S}}(W) - \mathscr{L}_S(W) - \mathbb{E}[\mathscr{L}_{\bar{S}}(W') - \mathscr{L}_S(W')],$$

where W' is an independent copy of W such that $P_{W',\tilde{S},U} = P_{W,\tilde{S}} \otimes P_U$. Hence, the inequalities $\mathbb{E}[\mathscr{L}_{\bar{S}}(W') - \mathscr{L}_S(W')] = 0$ and $|x+y| \leq |x| + |y|$ lead to the upper bound

$$\left|\mathbb{E}[\widehat{\mathrm{gen}}(W,\tilde{S},U)]\right| \leq \left|\mathbb{E}[\mathscr{L}_{\bar{S}}(W') - \mathscr{L}_{\bar{S}}(W)]\right| + \left|\mathbb{E}[\mathscr{L}_S(W') - \mathscr{L}_S(W)]\right|,$$

where the KR duality can be applied to each of the terms, albeit at the expense of an extra factor of 2.

3.3 Backward channel

In [17], the authors study the characterization of the expected generalization error in terms of the discrepancy between the data distribution P_S and the backward channel distribution $P_{S|W}$ motivated by its connection to rate–distortion theory, see e.g., [19, Chapters 25–57] or [22, Chapter 10]. An approach formalizing this intuitive connection is given in Appendix G and different angles, based on chaining mutual information [13] and compression, are found in [11, Section 5] and [23].

More concretely, they proved that the generalization error is bounded from above by the discrepancy of these distributions, where the discrepancy is measured by the Wasserstein distance of order p with the Minkowski distance of order p as a metric, i.e., $\rho(x, y) = \|x - y\|_p$. Namely,

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq \frac{L}{n^{1/p}} \mathbb{E}[\mathbb{W}^p_{p,|\cdot|}(P_S, P_{S|W})]^{1/p}.$$

Similarly, the results from §3.1 and §3.2 can be replicated considering the backward channel instead of the forward channel, e.g., $P_{S|W}$ instead of $P_{W|S}$ in (1), $P_{Z_i|W}$ instead of $P_{W|Z_i}$ in Theorem 1. However, in this case, the loss ℓ would be required to be Lipschitz with respect to the samples space $\mathcal{Z}$ and not the hypothesis space $\mathcal{W}$, i.e., Lipschitz for all fixed $w \in \mathcal{W}$, thus exploiting the geometry of the samples' space and not the hypotheses' one.

As an example, noting that $\overline{\mathrm{gen}}(W, S) = \mathbb{E}[\mathscr{L}_{S'}(W) - \mathscr{L}_S(W)]$, where S' is an independent copy of S such that $P_{W,S'} = P_W \otimes P_S$ produces the bound

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq L\mathbb{E}[\mathbb{W}(P_S, P_{S|W})].$$

Compared to [17], these results (i) are valid for any metric ρ as long as the loss ℓ is Lipschitz under ρ, and (ii) have single-letter and random-subset versions, and (iii) have variants in both the standard and randomized-subsample settings.

3.4 Other information measures

The bounds obtained in §3.1 and §3.2 may be manipulated to produce a variety of new bounds based on common information measures. For example, once the discrete metric is assumed and since the total variation is symmetric, applying Pinsker's inequality with the distributions in the opposite order to Corollaries 1, 2, 3, and 4 and further applying Jensen's inequality yields bounds based on the lautum information $\mathtt{L}$ [24]. For instance, a corollary of Theorem 3 is

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq \frac{b-a}{n} \sum\nolimits_{i=1}^{n} \Psi\big(\mathtt{L}(W; Z_i)\big).$$

Similarly, several new bounds based on different f*-divergences* [19, Chapter 7] may be obtained employing the *joint range strategy* once the discrete metric is assumed. As an example, some corollaries of Theorem 1 based on the Hellinger distance $\mathtt{H}$ and the χ^2-divergence (see Appendix H for a tighter and more general version of (6)) are

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq \frac{L}{2n} \sum\nolimits_{i=1}^{n} \mathbb{E}\left[\mathtt{H}(P_{W|Z_i}, P_W)\sqrt{4 - \mathtt{H}^2(P_{W|Z_i}, P_W)}\right], \tag{4}$$

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq \frac{L}{\sqrt{2}n} \sum\nolimits_{i=1}^{n} \mathbb{E}\left[\sqrt{\log\left(1 + \chi^2(P_{W|Z_i}, P_W)\right)}\right], \text{ and} \tag{5}$$

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq \frac{L}{2n} \sum\nolimits_{i=1}^{n} \mathbb{E}\left[\sqrt{\chi^2(P_{W|Z_i}, P_W)}\right]. \tag{6}$$

3.5 Final remarks on the generality of the results

Due to Bobkov–Götze's theorem [14, Theorem 4.8], the relative entropy results still hold when the loss is both Lipschitz and subgaussian. Hence, the presented Wasserstein distance bounds are tighter than [4, 6, 7, 9, 10] in a more general setting. Moreover, the total variation results also hold for any metric with the added factor of $d_\rho(\mathcal{W})$ as per Remark 3. These results were omitted in the main text for clarity of exposition, but are included in Appendix B.

Therefore, only when the loss is not Lipschitz but is subgaussian or has a bounded cumulant function, or $\mathcal{W}$ is not Polish, the bounds from [6, 7, 10, 12] are preferred. As an example, some common loss functions such as the cross-entropy, the Hinge loss, the Huber loss, or any L_p norm are Lipschitz [25, 26] under an appropriate metric ρ, see Appendix A for a discussion of the role of the metric and the space geometry in the presented bounds.

4 Discussion

This paper introduced several expected generalization error bounds based on the Wasserstein distance. In particular, these are full-dataset, single-letter, and random-subset bounds on both the standard and the randomized-subsample settings. When the Wasserstein distance ignores the geometry of the hypothesis space and the loss is bounded, the presented bounds are tighter and recover from below the current bounds based on the relative entropy and the mutual information [4, 6, 7, 9, 10], see also Appendix B for stronger, more general statements. Furthermore, the obtained total variation and relative-entropy bounds on the standard setting are ensured to be non-vacuous, i.e., smaller or equal than the trivial bound, thus resolving the issue of potentially vacuous relative-entropy and mutual-information bounds on the standard setting. Interestingly, the results for the randomized-subsample setting are tighter than their analogous in the standard setting only if their Wasserstein distance (or total variation) is twice as small.

Moreover, the techniques employed to obtain these bounds can also be used to obtain analogous bounds considering the backward channel and the samples' space geometry, aiming to facilitate connections between the generalization error characterization and rate–distortion theory, as suggested by Lopez and Jog [17]. Nonetheless, when the backward channel can be characterized, these bounds are interesting in their own right. Finally, the presented bounds may be used to generate a variety of new bounds in terms of, e.g., the lautum information or f-divergences like the total variation, the relative entropy, the Hellinger distance, or the χ^2-divergence.

4.1 Limitations and future work

PAC-Bayes bounds PAC-Bayes bounds ensure that $\mathbb{E}[\operatorname{gen}(W,S) \mid S] \geq \alpha(\beta^{-1})$ with probability no greater than $\beta \in (0,1)$. Similarly, single-draw PAC-Bayes bounds ensure that $\operatorname{gen}(W,S) \geq \alpha(\beta^{-1})$ with probability no greater than $\beta \in (0,1)$. These concentration bounds are of high probability when the dependency on β^{-1} is logarithmic, i.e., $\log(1/\beta)$. See, [27, 2] for an overview.

The bounds from this work may be used to obtain single-draw PAC-Bayes bounds applying Markov's inequality [22, Problem 3.1] directly. For instance, employing it in Theorem 1 implies that

$$P_{W,S}\Big(\operatorname{gen}(W,S) \geq \frac{L}{\beta n}\sum\nolimits_{i=1}^{n}\mathbb{E}[\mathbb{W}(P_{W|Z_i},P_W)]\Big) \leq \beta,$$

for all $\beta \in (0,1)$. However, this is not a high-probability bound since the dependency on β^{-1} is linear. Hence, high-probability concentration bounds based on the Wasserstein distance and the total variation are a path of future research. As an example, [28–31] provide high-probability single-draw PAC-Bayes bounds based on, respectively, max-information, differential privacy, α-mutual information, and uniform stability. Similarly, high-probability PAC-Bayes bounds based on the relative entropy and the hypothesis' space geometry are given in [8] and [32, 33], respectively.

New bounds to specific algorithms The Wasserstein distance is difficult to characterize and/or estimate. Nonetheless, some of the bounds that can be obtained from it, e.g., mutual-information and relative-entropy bounds, have been used to obtain analytical bounds on specific algorithms, e.g., Langevin dynamics and stochastic gradient Langevin dynamics [6, 7, 9, 10]. Some of these results can be readily tightened with Corollaries 1, 2, and 4. Thence, deriving new analytical bounds for specific algorithms based on the presented results is also a topic for further research.

Connections to stability and privacy measures A learning algorithm is said to be stable if a small change on the input dataset produces a small variation in the output hypothesis. There are various attempts at quantifying this notion such as uniform stability [34], where the variation in the output hypothesis is seen in terms of the loss, and differential privacy (DP) [28], where this variation is seen in terms of the hypothesis distribution. These notions are tied to the generalization capability of an algorithm, i.e., the less a hypothesis depends on the specifics of the data samples, the better it will generalize, and hence there are works obtaining generalization bounds based on stability, see e.g., [29, 31]. In particular, there are some works that, assuming some stability notion such as DP, bound from above the relative entropy and the mutual information appearing in some of the bounds that can be derived from the results presented in this work, hence also tying stability and generalization, c.f. [1, 35, 36]. Therefore, a future line of research is to investigate how different notions of stability can be combined with the measures of similarity between distributions employed in this work to characterize the generalization error.

Funding

This work was funded in part by the Swedish research council under contract 2019-03606.

References

[1] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 125, Jul. 2020, pp. 3437–3452.

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

[3] V. Vapnik, *The Nature of Statistical Learning Theory*, ser. Information Science and Statistics. Springer Science & Business Media, 2013.

[4] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, 2017, pp. 2524–2533.

[5] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, Jan. 2020.

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

[7] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for SGLD via data-dependent estimates," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 015–11 025.

[8] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 824–839, Nov. 2020.

[9] B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm," in *IEEE Information Theory Workshop (ITW)*. IEEE, 2020.

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

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

[12] R. Zhou, C. Tian, and T. Liu, "Individually conditional individual mutual information bound on generalization error," *arXiv preprint arXiv:2012.09922*, 2020.

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

[14] R. van Handel, "Probability in high dimension," Princeton University, NJ, Tech. Rep., 2014.

[15] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via wasserstein distance," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 577–581.

[16] J. Zhang, T. Liu, and D. Tao, "An optimal transport view on generalization," *arXiv preprint arXiv:1811.03270*, 2018.

[17] A. T. Lopez and V. Jog, "Generalization error bounds using wasserstein distances," in *2018 IEEE Information Theory Workshop (ITW)*. IEEE, 2018, pp. 1–5.

[18] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer Science & Business Media, 2008, vol. 338.

[19] Y. Polyanskiy and Y. Wu, "Lecture notes on Information Theory," *MIT (6.441), UIUC (ECE 563), Yale (STAT 664)*, 2017.

[20] J. Bretagnolle and C. Huber, "Estimation des densités: risque minimax," in *Séminaire de Probabilités XII*. Springer, 1978, pp. 342–363, in French.

[21] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *2016 IEEE Information Theory Workshop (ITW)*. IEEE, 2016, pp. 26–30.

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[23] Y. Bu, W. Gao, S. Zou, and V. Veeravalli, "Information-theoretic understanding of population risk improvement with model compression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 3300–3307, Apr. 2020.

[24] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, 2008.

[25] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[26] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," 2018.

[27] B. Guedj, "A primer on PAC-Bayesian learning," *arXiv preprint arXiv:1901.05353*, 2019.

[28] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[29] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Advances in Neural Information Processing Systems*, 2015, pp. 2350–2358.

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

[31] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy, "Sharper bounds for uniformly stable algorithms," in *Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, 2020, pp. 610–626.

[32] J.-Y. Audibert and O. Bousquet, "Pac-bayesian generic chaining." in *NIPS*. Citeseer, 2003, pp. 1125–1132.

[33] J.-Y. Audibert and O. Bousquert, "Combining pac-bayesian and generic chaining bounds." *Journal of Machine Learning Research*, vol. 8, no. 4, 2007.

[34] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of machine learning research*, vol. 2, no. Mar, pp. 499–526, 2002.

[35] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conference*. Springer, 2016, pp. 635–658.

[36] B. Rodríguez-Gálvez, G. Bassi, and M. Skoglund, "Upper bounds on the generalization error of private algorithms for discrete data," *IEEE Transactions on Information Theory*, vol. 67, no. 11, pp. 7362–7379, 2021.

[37] D. Haussler, "Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension," *Journal of Combinatorial Theory, Series A*, vol. 69, no. 2, pp. 217–232, 1995.

[38] V. Feldman and T. Steinke, "Calibrating noise to variance in adaptive data analysis," pp. 535–544, 2018. [Online]. Available: http://proceedings.mlr.press/v75/feldman18a.html

[39] J. N. McDonald and N. A. Weiss, *A Course in Real Analysis*, 2nd ed. Cambridge, Massachusetts: Elsevier, 2013.

[40] R. M. Gray, *Source coding theory*, ser. Engineering and Computer Science. Springer Science & Business Media, 2012, vol. 83.

[41] Y. Wu, "Lecture notes on information-theoretic methods for high-dimensional statistics," *Lecture Notes for ECE598YW (UIUC)*, vol. 16, 2017.

[42] T. Popoviciu, "Sur les équations algébriques ayant toutes leurs racines réelles," *Mathematica*, vol. 9, pp. 129–145, 1935.

[43] P. Harremoës and I. Vajda, "On pairs of f-divergences and their joint range," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3230–3235, 2011.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Section 3 follows the abstract statements in order.
   (b) Did you describe the limitations of your work? [Yes] See Section 4.1.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We are unaware of potential negative societal impacts of this line of work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The i.i.d. assumption is in the first paragraph of the introduction. All other assumptions are in the statement of all theorems and corollaries.
   (b) Did you include complete proofs of all theoretical results? [Yes] All theoretical results are proved in completion in the appendix. Nonetheless, some intuition and the most important parts are also given in the main text.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL?[N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A A short note on geometry and generalization error bounds

A.1 Chaining- and Wasserstein-based bounds

Ever since it was shown that some infinitely-dimensional hypothesis spaces where PAC-learnable thanks to the VC dimension [2, Chapter 6], there has been an interest in studying the role of the space complexity and its geometry in determining the generalization of an algorithm, e.g. [37].

Some of the most interesting results come from the theory of random processes. Adapted to our notation, this theory considers the set of random variables (or random process) $\{\text{gen}(w,S)\}_{w\in\mathcal{W}}$. Then, this theory bounds the generalization error using the ϵ-covering number of the hypothesis space $\mathcal{N}(\mathcal{W},\rho,\epsilon)$ with some requirements on the smoothness of the hypothesis' space $\mathcal{W}$ under a metric ρ. Here, the geometry of the space captures its complexity via the ϵ-covering number, which is defined as the cardinality of the minimum set $\mathcal{N}$ such that for all hypothesis w in $\mathcal{W}$, there is an element of the set $x\in\mathcal{N}$ such that $\rho(x,w)\leq\epsilon$. The two main techniques from this line of work are:

- The Lipschitz maximal inequality [14, Lemma 5.7]. Here, the smoothness condition is to require the process to be Lipschitz; that is, that there is a random variable C such that for all $w,w'\in\mathcal{W}$
$$|\text{gen}(w,S)-\text{gen}(w',S)|\leq C\rho(w,w').$$
Then, with the additional condition that $\text{gen}(w,S)$ should be σ-subgaussian under P_S for all $w\in\mathcal{W}$, this inequality bounds the generalization error as follows:
$$\mathbb{E}[\text{gen}(W,S)]\leq\inf_{\epsilon\in\mathbb{R}}\left\{\epsilon\mathbb{E}[C]+\sqrt{2\sigma^2\log\mathcal{N}(\mathcal{W},\rho,\epsilon)}\right\}.$$
Note how this technique creates a tension between the finesse ϵ of the covering and the smoothness of the process $\mathbb{E}[C]$.
- Dudley's chaining technique [14, Theorem 5.24]. Here, the smoothness condition is to require the process to be subgaussian; that is, for all $w,w'\in\mathcal{W}$ and all $\lambda\geq 0$,
$$\log\mathbb{E}\left[e^{\lambda\left(\text{gen}(w,S)-\text{gen}(w',S)\right)}\right]\leq\frac{\lambda^2\rho\left(w,w'\right)^2}{2}$$
and $\mathbb{E}[\text{gen}(w,S)]=0$. Then, with the additional condition that the process is separable, this inequality bounds the generalization error as follows:
$$\mathbb{E}[\text{gen}(W,S)]\leq 6\sum_{k\in\mathbb{Z}}2^{-k}\sqrt{\log\mathcal{N}(\mathcal{W},\rho,2^{-k})}.$$
Note that the subgaussian requirement is a relaxation of the Lipschitz requirement. As such, now the bound is expressed as a sum where each finer covering number of the space is weighted by the finesse (2^{-k}) of such a covering. Further refined bounds based on this technique can be found in [32, 33].

The first work that included the relationship between the hypothesis and the training samples to the analysis using random processes was [13]. There, the authors combined the chaining technique with [4, Lemma 1] to derive a formula which bounds the generalization error by a weighted average of the mutual information between the dataset and increasingly finer quantizations W_k of the hypothesis. More precisely, they proved that
$$\mathbb{E}[\text{gen}(W,S)]\leq 3\sqrt{2}\sum_{k\in\mathbb{Z}}2^{-k}\sqrt{I(W_k;S)}.$$
Therefore, in [13], the geometry of the hypothesis space $\mathcal{W}$ under the metric ρ is expressed as the amount of information that the quantizations of the hypothesis under ρ contain about the dataset S.

On the other hand, in [15] and this paper, a stronger smoothness requirement is considered. Namely that the loss function is L-Lipschitz; i.e., $|\ell(w,z)-\ell(w',z)|\leq L\rho(w,w')$ for all $w,w'\in\mathcal{W}$ and all $z\in\mathcal{Z}$. This is a stronger statement since the subgaussian assumption can be viewed as an "in-probability" version of the Lipschitz assumption. Nonetheless, this stronger assumption allows these bounds to bound the generalization error by the minimum cost to go from the distribution of the hypothesis after observing some samples to its marginal distribution, e.g., $\mathbb{W}(P_{W|Z_i},P_W)$. Here the

metric ρ is used to quantify how far away are the realizations of both probability distributions, on average, if they are coupled in the best way possible.

An interesting property of the presented bounds is that they have the flexibility to consider either the hypothesis or the sample space (backward channel). For instance, for [13, Example 1], the presented bounds using the backward channel are tighter than the bounds arising from the chaining technique.

In this example, the authors consider the canonical Gaussian process. The hypothesis space is $\mathcal{W} = \{w \in \mathbb{R}^2 : \|w\|_2 = 1\}$, the samples are $Z_i \sim \mathcal{N}(0, I_2)$, the loss function is $\ell(w, z) = -w^T z$, and the hypothesis is selected with the empirical risk minimization (ERM) algorithm, i.e., $w^\star = \arg\min_{w \in \mathcal{W}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(w, z_i) \right\}$. In this setting, by Cauchy–Schwarz we see that the loss $\ell(w, \cdot)$ is 1-Lipschitz for all $w \in \mathcal{W}$; i.e., $|-w^T z + w^T z'| \leq \|w\|_2 \|z - z'\|_2 \leq \|z - z'\|_2$ for all $z, z' \in \mathcal{Z}$. Therefore, the backward channel equivalent of Theorem 1 holds. Moreover, the function is 1-subgaussian under P_Z for all $w \in \mathcal{W}$. Therefore, as shown in Appendix B.1 the presented bound is tighter than [23], which is shown to be tighter than [13] in this setting (see [23, Section IV-B]).

A.2 Choice of the metric

The presented bounds in Sections 3.1, 3.2, and 3.3 are valid for any metric ρ under which the loss is Lipschitz. The Lipschitz property is a property of the hypothesis space $\mathcal{W}$ (forward channel bounds) or the sample space $\mathcal{Z}$ (backward channel bounds) and *not* of the algorithm. That is, if two different algorithms operate on the same sample space and produce the same hypothesis space, then they can be characterized with the same metric.

The choice of the metric can be decisive for a tight analysis of the presented bounds, and there are times where a loss function can be Lipschitz under several metrics. For example, a bounded loss function represented as a norm is Lipschitz with respect to that norm and the discrete metric. Nonetheless, in many situations the metric of choice becomes apparent based on the loss function. For example, if we consider the forward channel bounds and samples of the type $z = (x, y)$ and the following two common supervised tasks:

- Regression. If a norm is used as the loss function $\ell(w, z) = \|w - y\|$, then such a norm is also a good choice for a metric since by the reverse triangle inequality the loss is 1-Lipschitz under that metric: $\big|\|w - y\| - \|w' - y\|\big| \leq \|w - w'\|$ for all $w, w' \in \mathcal{W}$.
- Classification. If the 0-1 loss is used as the loss function $\ell(w, z) = \mathrm{Ind}(w \neq y)$, then the discrete metric is a good choice since the loss is also 1-Lipschitz under this metric: $|\mathrm{Ind}(w \neq y) - \mathrm{Ind}(w' \neq y)| \leq \mathrm{Ind}(w \neq w')$.

Similarly, for the backward channel bounds, it is known that the logistic loss, the softmax loss,[2] the Hinge loss, and many distance-based losses like norms, the Huber, ϵ-insensitive, and pinball losses, are Lipschitz under the L_1 norm metric $\rho(z, z') = |z - z'|$ [25, 26].

B Generality of the results

The main text only presents total variation and relative entropy bounds for bounded losses. This is to clarify the (geometrical) relationship between the bounds based on the Wasserstein distance and those based on the relative entropy. That is, that the former recover the latter when the geometry of the hypothesis space is ignored.

Nonetheless, these bounds hold more generally for any metric under certain mild modifications. More concretely, the relative entropy bounds hold when the loss is also subgaussian[3]. Furthermore, at the cost of an extra factor of $d_\rho(\mathcal{W})$, the two kinds of bound still hold.

B.1 Extension of the relative entropy bounds to subgaussian losses

Consider a Polish space $(\mathcal{X}, \rho)$ and a probability distribution P on $\mathcal{X}$ with a finite first moment. Then, the Bobkov–Götze's theorem [14, Theorem 4.8] says that the following statements are equivalent:

[2] This result can be derived from [26, Proposition 3] and the $L_1 - L_2$ inequality.

[3] A random variable X is said to be σ-subgaussian if $\log \mathbb{E}[\exp \lambda(X - \mathbb{E}[X])] \leq \frac{\lambda^2 \sigma^2}{2}$ for all $\lambda \in \mathbb{R}$. Also, a function $f : \mathcal{X} \to \mathbb{R}$ is said to be σ-subgaussian under P if $f(X)$ is σ-subgaussian and $X \sim P$.

- f is σ-subgaussian under P *for every* 1-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$; and,
- $\mathbb{W}(Q, P) \leq \sqrt{2\sigma^2 D_{\mathrm{KL}}(Q \,\|\, P)}$ for all Q on $\mathcal{X}$.

Therefore, the results from Corollaries 1, 2, 3, and 4 are valid in a more general setting, namely when the loss function ℓ is both L-Lipschitz and σ-subgaussian for all $z \in \mathcal{Z}$. To realize this, note that if $X \sim P$ is L-Lipschitz and σ-subgaussian, then X/L is 1-Lipschitz and (σ/L)-subgaussian. Hence, if $|X - \mathbb{E}[X]| \leq L\mathbb{W}(Q, P)$, then it follows that $|X - \mathbb{E}[X]| \leq L\sqrt{2(\sigma/L)^2 D_{\mathrm{KL}}(Q \,\|\, P)} = \sqrt{2\sigma^2 D_{\mathrm{KL}}(Q \,\|\, P)}$.

As an example, assume that the loss function ℓ is L-Lipschitz and σ-subgaussian under P_W for all $z \in \mathcal{Z}$ and that $\mathcal{W}$ is Polish. Then,

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq \frac{L}{n}\sum_{i=1}^{n} \mathbb{E}\big[\mathbb{W}(P_{W|Z_i}, P_W)\big] \leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2\sigma^2 I(W; Z_i)} \leq \sqrt{\frac{2\sigma^2 I(W; S)}{n}}$$

is a corollary of Theorem 1 due to Bobkov–Götze's theorem. As when the loss is bounded, this equation shows that Theorem 1 is tighter than [6, Proposition 2] and [4, Theorem 1] and exhibits a decaying factor of $1/\sqrt{n}$. Moreover, this result encompasses the case where the loss is bounded in $[a, b]$ since if a random variable is bounded in $[a, b]$ it is $(b - a)/2$-subgaussian.

Note, however, that in this case the subgaussianity constant is different than the constant from, e.g. [4], where the loss ℓ was supposed to be ν-subgaussian under P_Z for all $w \in \mathcal{W}$. Considering the bounds based on the backward channel (§3.3), these constants are exactly the same, since assuming that the loss function ℓ is L-Lipschitz and ν-subgaussian under P_Z for all $w \in \mathcal{W}$ and that $\mathcal{Z}$ is Polish, means that

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq \frac{L}{n}\sum_{i=1}^{n} \mathbb{E}\big[\mathbb{W}(P_{Z_i|W}, P_{Z_i})\big] \leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2\nu^2 I(W; Z_i)} \leq \sqrt{\frac{2\nu^2 I(W; S)}{n}},$$

which is always tighter than [4, 6]. As mentioned above, when the loss is bounded the subgaussianity constant is the same under any distribution.

B.2 Extension to the total variation bounds for any metric

As per Remark 3, the results based on the total variation also hold for any metric with the extra factor $d_\rho(\mathcal{W})$, where $d_\rho(\mathcal{W})$ is the diameter of the hypothesis space $\mathcal{W}$ under the metric ρ. For instance, Corollary 1 results in

$$\left|\overline{\mathrm{gen}}(W, S)\right| \leq \frac{L d_\rho(\mathcal{W})}{n}\sum_{i=1}^{n} \mathbb{E}\big[\mathtt{TV}(P_{W|Z_i}, P_W)\big] \leq \frac{L d_\rho(\mathcal{W})}{n}\sum_{i=1}^{n} \mathbb{E}\big[\Psi\big(D_{\mathrm{KL}}(P_{W|Z_i} \,\|\, P_W)\big)\big].$$

Note that, since the results based on the total variation hold for any metric, all the derived results based on different information measures from §3.4 hold too.

The extra term can be arbitrarily large as, for example, when the hypothesis is the weights of a neural network and metric ρ is the ℓ_2 norm $\|\cdot\|_2$. Nonetheless, this term can still be small and relevant for practical settings. For instance, consider again that the hypothesis is the weights of a neural network. However, consider now that the metric ρ is the infinity norm $\|\cdot\|_\infty$ and that each weight is enforced to be smaller than some small constant C. Then, the diameter of the space $d_{\|\cdot\|_\infty}(\mathcal{W})$ is (at most) equal to C.

C Proofs of the theorems from Section 3

C.1 Proof of Theorem 1

Note that $\mathbb{E}[\mathscr{L}_{P_Z}(W)] = \int_{\mathcal{W}\times\mathcal{Z}} \ell(w,z)d(P_W \otimes P_Z)(w,z)$. If W' is an independent copy of W such that $P_{W',Z_i} = P_W \otimes P_Z$ for all $i \in [n]$, then

$$\begin{aligned}\left|\overline{\text{gen}}(W,S)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\ell(W',Z_i) - \ell(W,Z_i)\big]\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\mathbb{E}\big[\ell(W',Z_i) - \ell(W,Z_i) \mid Z_i\big]\Big]\right| \\ &\leq \frac{L}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathbb{W}(P_{W|Z_i}, P_W)\big],\end{aligned}$$

where the last inequality stems from the KR duality and the Lipschitzness of ℓ for all $z \in \mathcal{Z}$. The absolute value is removed since ρ is a metric.

C.2 Proof of Theorem 2

Consider the quantity

$$\text{gen}_j(w, s_j) \triangleq \mathscr{L}_{P_Z}(w) - \mathscr{L}_{s_j}(w),$$

where $\mathscr{L}_{s_j} = \frac{1}{m}\sum_{i\in j}\ell(w,z_i)$. Then, note that $\mathbb{E}[\text{gen}(W,S)] = \mathbb{E}[\text{gen}_J(W,S_J)]$ if $\mathscr{L}_s(w) = \mathbb{E}[\mathscr{L}_{s_J}(w)]$. This last equality follows since J is uniformly distributed, there are $\binom{n}{m}$ possible subsets of size m in $[n]$, and each sample z_i belongs to $\binom{n-1}{m-1}$ of those subsets; hence

$$\mathbb{E}[\mathscr{L}_{s_J}(w)] = \frac{1}{\binom{n}{m}}\sum_{j\in\mathcal{J}}\frac{1}{m}\sum_{i\in j}\ell(w,z_i) = \frac{1}{n}\sum_{i=1}^{n}\ell(w,z_i) = \mathscr{L}_s(w). \tag{7}$$

Subsequently, the two bounds from the theorem are obtained after bounding the innermost expectation on the right hand side of the following equation:

$$\left|\overline{\text{gen}}(W,S)\right| = \left|\mathbb{E}[\text{gen}_J(W,S_J)\right| \leq \mathbb{E}\big[\big|\mathbb{E}[\text{gen}_J(W,S_J) \mid J, S_{J^c}, R]\big|\big],$$

where the last step follows from Jensen's inequality.

(a) Consider, until stated otherwise, that all random objects and expectations are conditioned to a fixed j, s_{j^c}, and r. Then note that $\mathbb{E}[\mathscr{L}_{P_Z}(W)] = \mathbb{E}[\mathscr{L}_{S_j}(W')]$, where W' is an independent copy of W such that $P_{W',S_j|s_{j^c},r} = P_{W|s_{j^c},r} \otimes P_{S_j}$. Therefore

$$\left|\mathbb{E}[\text{gen}_j(W,S_j)]\right| \leq \left|\mathbb{E}[\mathscr{L}_{S_j}(W') - \mathscr{L}_{S_j}(W)]\right| \leq L\mathbb{E}[\mathbb{W}(P_{W|S_j,s_{j^c},r}, P_{W|s_{j^c},r})],$$

where the last inequality stems from the KR duality. Finally, taking the expectation with respect to $P_{J,S_{J^c},R}$ in both sides of the equation completes the proof.

(b) Consider again, until stated otherwise, that all random objects and expectations are conditioned to a fixed j, s_{j^c}, and r. Then, similarly to the proof of Theorem 1

$$\begin{aligned}\left|\mathbb{E}[\text{gen}_j(W,S_j)]\right| &\leq \left|\frac{1}{m}\sum_{i\in j}\mathbb{E}[\ell(W',Z_i) - \ell(W,Z_i)]\right| \\ &\leq \frac{L}{m}\sum_{i\in j}\mathbb{E}[\mathbb{W}(P_{W|Z_i,s_{j^c},r}, P_{W|s_{j^c},r})],\end{aligned}$$

where the last inequality stems from the KR duality and W' is an independent copy of W such that $P_{W',Z_i|s_{j^c},r} = P_{W'|s_{j^c},r} \otimes P_{Z_i}$ for all $i \in j$. Finally, taking the expectation with respect to $P_{J,S_{J^c},R}$ in both sides of the equation completes the proof.

C.3 Proof of Equation 3

Consider an independent copy W' of W such that $P_{W',\tilde{S},U} = P_{W',\tilde{S}} \otimes P_U$. Then, $\mathbb{E}[\mathscr{L}_S(W')] = \mathbb{E}[\mathscr{L}_{\bar{S}}(W')]$, and therefore

$$\widehat{\text{gen}}(w, \tilde{s}, u) = \mathscr{L}_{\bar{s}}(w) - \mathscr{L}_s(w) - \mathbb{E}[\mathscr{L}_{\bar{S}}(W') - \mathscr{L}_S(W')].$$

Then, re-arranging the expectation of the above expression and using the fact that $|x+y| \leq |x| + |y|$ results in

$$\begin{aligned}
\left|\mathbb{E}[\widehat{\text{gen}}(W, \tilde{S}, U)]\right| &\leq \left|\mathbb{E}[\mathscr{L}_{\bar{S}}(W') - \mathscr{L}_{\bar{S}}(W)]\right| + \left|\mathbb{E}[\mathscr{L}_S(W') - \mathscr{L}_S(W)]\right| \\
&= \left|\mathbb{E}\left[\mathbb{E}[\mathscr{L}_{\bar{S}}(W') - \mathscr{L}_{\bar{S}}(W) \mid \tilde{S}, U]\right]\right| + \left|\mathbb{E}\left[\mathbb{E}[\mathscr{L}_S(W') - \mathscr{L}_S(W) \mid \tilde{S}, U]\right]\right| \\
&\leq 2L\mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})],
\end{aligned}$$

where the last step stems from the KR duality. Finally, noting that $\mathbb{E}[\widehat{\text{gen}}(W, \tilde{S}, U)] = \overline{\text{gen}}(W, S)$ completes the proof.

C.4 Proof of Theorem 3

Similarly to the proof of (1), consider an independent copy W' of W such that $P_{W',\tilde{S}_i,U_i} = P_{W',\tilde{S}_i} \otimes P_{U_i}$. Then $\mathbb{E}[\ell(W', \bar{Z}_i)] = \mathbb{E}[\ell(W', Z_i)]$, where $Z_i = \tilde{Z}_{i+U_i n}$, $\bar{Z}_i = \tilde{Z}_{i+(1-U_i)n}$, and $\tilde{S}_i = (Z_i, \bar{Z}_i)$. Therefore,

$$\widehat{\text{gen}}(w, \tilde{s}, u) = \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(w, \bar{z}_i) - \ell(w, z_i) - \mathbb{E}[\ell(W', \bar{Z}_i) - \ell(W', Z_i)]\Big).$$

Then, re-arranging the expectation of the above expression and using the fact that $\left|\sum_{i=1}^n x_i\right| \leq \sum_{i=1}^n |x_i|$ results in

$$\begin{aligned}
\left|\widehat{\text{gen}}(W, \tilde{S}, U)\right| &\leq \frac{1}{n}\sum_{i=1}^{n}\Big(\left|\mathbb{E}[\ell(W', \bar{Z}_i) - \ell(W, \bar{Z}_i)]\right| + \left|\mathbb{E}[\ell(W', Z_i) - \ell(W, Z_i)]\right|\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\left|\mathbb{E}\left[\mathbb{E}[\ell(W', \bar{Z}_i) - \ell(W, \bar{Z}_i) \mid \tilde{S}_i, U_i]\right]\right| + \left|\mathbb{E}\left[\mathbb{E}[\ell(W', Z_i) - \ell(W, Z_i) \mid \tilde{S}_i, U_i]\right]\right|\Big) \\
&\leq \frac{2L}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}_i})\right],
\end{aligned}$$

where the last step stems from the KR duality. Finally, noting that $\mathbb{E}[\widehat{\text{gen}}(W, \tilde{S}, U)] = \overline{\text{gen}}(W, S)$ completes the proof.

C.5 Proof of Theorem 4

Similarly to the proof of Theorem 2, consider the quantity

$$\widehat{\text{gen}}_j(w, \tilde{s}_j, u_j) \triangleq \mathscr{L}_{\bar{s}_j}(w) - \mathscr{L}_{s_j}(w),$$

where $\tilde{s}_j = (s_j, \bar{s}_j)$ and s_j and $\bar{s}_j$ are the subsets of s and $\bar{s}$, respectively, indexed by j. Then, note that $\mathbb{E}[\widehat{\text{gen}}(W, \tilde{S}, U)] = \mathbb{E}[\widehat{\text{gen}}_J(W, \tilde{S}_J, U_J)]$ since, as shown in (7), the equalities $\mathbb{E}[\mathscr{L}_{s_j}(w)] = \mathscr{L}_s(w)$ and $\mathbb{E}[\mathscr{L}_{\bar{s}_j}(w)] = \mathscr{L}_{\bar{s}}(w)$ hold.

Then, the two bounds from the theorem are obtained after bounding the innermost expectation on the right hand side of the following equation:

$$\left|\mathbb{E}[\widehat{\text{gen}}(W, \tilde{S}, U)]\right| = \left|\mathbb{E}[\widehat{\text{gen}}_J(W, \tilde{S}_J, U_J)]\right| \leq \mathbb{E}\left[\left|\mathbb{E}[\widehat{\text{gen}}_J(W, \tilde{S}_J, U_J) \mid J, \tilde{S}, U_{J^c}, R]\right|\right],$$

where the last step follows from Jensen's inequality.

(a) Consider, until stated otherwise, that all random objects and expectations are conditioned to a fixed j, $\tilde{s}$, u_{j^c}, and r. Then note that $\mathbb{E}[\mathscr{L}_{\bar{s}_j}(W')] = \mathbb{E}[\mathscr{L}_{s_j}(W')]$, where W' is an independent copy of W such that $P_{W',U_j|\tilde{s},u_{j^c},r} = P_{W'|\tilde{s},u_{j^c},r} \otimes P_{U_j}$. Therefore,

$$\widehat{\text{gen}}_j(w, \tilde{s}_j, u_j) = \mathscr{L}_{\bar{s}_j}(w) - \mathscr{L}_{s_j}(w) - \mathbb{E}[\mathscr{L}_{\bar{s}_j}(W') - \mathscr{L}_{s_j}(W')].$$

Then, re-arranging the expectation of the above expression and using the fact that $|x+y| \leq |x| + |y|$ results in

$$\begin{aligned}\left|\widehat{\text{gen}}_j(W, \tilde{s}_j, U_j)\right| &\leq \left|\mathbb{E}[\mathscr{L}_{\bar{s}_j}(W') - \mathscr{L}_{\bar{s}_j}(W)]\right| + \left|\mathbb{E}[\mathscr{L}_{s_j}(W') - \mathscr{L}_{s_j}(W)]\right| \\ &= \left|\mathbb{E}\big[\mathbb{E}[\mathscr{L}_{\bar{s}_j}(W') - \mathscr{L}_{\bar{s}_j}(W) \mid U_j]\big]\right| + \left|\mathbb{E}\big[\mathbb{E}[\mathscr{L}_{s_j}(W') - \mathscr{L}_{s_j}(W) \mid U_j]\big]\right| \\ &\leq 2L\mathbb{E}[\mathbb{W}(P_{W|U_j,\tilde{s},u_{j^c},r}, P_{W|\tilde{s},u_{j^c},r})],\end{aligned}$$

where the last inequality stems from the KR duality. Finally, taking the expectation with respect to $P_{J,\tilde{S},U_{J^c},R}$ in both sides of the equation completes the proof.

(b) Consider again, until stated otherwise, that all random objects and expectations are conditioned to a a fixed j, $\tilde{s}$, u_{j^c}, and r. Then, similarly to the proof of Theorem 3

$$\begin{aligned}\left|\widehat{\text{gen}}_j(W, \tilde{s}, U_j)\right| &\leq \frac{1}{m}\sum_{i\in j}\Big(\left|\mathbb{E}[\ell(W', \bar{Z}_i) - \ell(W, \bar{Z}_i)]\right| + \left|\mathbb{E}[\ell(W', Z_i) - \ell(W, Z_i)]\right|\Big) \\ &= \frac{1}{m}\sum_{i\in j}\Big(\left|\mathbb{E}\big[\mathbb{E}[\ell(W', \bar{Z}_i) - \ell(W, \bar{Z}_i) \mid U_i]\big]\right| + \left|\mathbb{E}\big[\mathbb{E}[\ell(W', Z_i) - \ell(W, Z_i) \mid U_i]\big]\right|\Big) \\ &\leq \frac{2L}{m}\sum_{i\in j}\mathbb{E}\big[\mathbb{W}(P_{W|U_i,\tilde{s},u_{j^c},r}, P_{W|\tilde{s},u_{j^c},r})\big],\end{aligned}$$

where the last inequality stems from the KR duality. Finally, taking the expectation with respect to $P_{J,\tilde{S},U_{J^c},R}$ in both sides of the equation completes the proof.

D Comparison of the bounds

In this section of the appendix, the full-dataset, single-letter, and random-subset bounds are compared. First, remember that the bounds based on the Wasserstein distance are tighter than the respective ones based on the relative entropy, and thus, those based on the mutual information under the conditions specified in Appendix B. The comparison in terms of the Wasserstein distance and in terms of the mutual information are found in §D.1 and in D.2, respectively.

When the bounds are compared in terms of the mutual information and in terms of the Wasserstein distance, the individual-sample bounds are shown to be tighter than the full-dataset bounds. Therefore, the idea that the individual forward channels $P_{W|Z_i}$, which are smoothed versions of the full-dataset forward channel $P_{W|S}$, are closer to the hypothesis marginal distribution P_W is backed by the theory in both cases.

Then, both cases also agree in that individual-sample bounds are tighter than random-subset bounds. Even though the proof for the Wasserstein distance based bounds also follows from a smoothing argument, a better insight is gained through the proof for the mutual information. These are based on the fact that $I(W;Z_i) \leq I(W;Z_i|S_{\mathcal{A}})$, where $\mathcal{A}$ is a subset of $[n]$ where i is not included. This inequality holds since the knowledge of the samples $S_{\mathcal{A}}$ provides information about Z_i through W. More precisely,

$$\begin{aligned}I(W;Z_i) &\overset{(a)}{\leq} I(W;Z_i) + \overbrace{I(S_{\mathcal{A}};Z_i|W)}^{\text{extra information}} \\ &\overset{(b)}{=} I(W, S_{\mathcal{A}};Z_i) \\ &\overset{(c)}{=} I(W;Z_i|S_{\mathcal{A}}) + I(Z_i;S_{\mathcal{A}}) \\ &\overset{(d)}{=} I(W;Z_i|S_{\mathcal{A}}),\end{aligned}$$

where (a) is due to the non-negativity of the mutual information, (b) and (c) follow from the chain rule, and (d) stems from the fact that Z_i and $S_{\mathcal{A}}$ are independent.

Finally, the comparison of the random-subset and full-dataset bounds behaves differently when the Wasserstein distance or the mutual information is employed. The analysis using the Wasserstein distance suggests that the random-subset bounds are sharper than the full-dataset bounds with an extra factor of two, namely

$$\mathbb{E}[\mathbb{W}(P_{W|S}, P_{W|S_{J^c}})] \leq 2\mathbb{E}[\mathbb{W}(P_{W|S}, P_W)].$$

On the other hand, the analysis using the mutual information indicates that the random-subset bounds are looser than the full dataset bounds.

This discrepancy might help understand the reason why, in practice, the random-subset bounds from [7, 10, 9] with the expectation of the square root of the relative entropy result in tighter characterizations of the generalization error than the full-dataset bounds from [4, 1] using the mutual information. To be precise, this suggests that the loss of performance of a further application of the Jensen's inequality to incorporate the expectations inside the square root of the bounds from [7, 10, 9] is big. In other words, that $D_{\mathrm{KL}}(P_{W|S} \,\|\, P_{W|S_J})$ has high variance in practical settings.

A summary of these comparisons is shown in Figure 2. Note that the mutual information bounds are written after Jensen's inequality is applied in order to allow comparisons between them. However, the relationships between the Wasserstein and mutual information based bounds still hold when the expectations of the relative entropy are outside of the square root.

D.1 Comparison of the Wasserstein distance based bounds

D.1.1 Standard setting

Proposition 2. *Consider the standard setting. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\mathbb{W}(P_{W|Z_i}, P_W)\big] \leq \mathbb{E}\big[\mathbb{W}(P_{W|S}, P_W)\big].$$

Proof. The proposition follows by noting that, for all $i \in [n]$,

$$\mathbb{E}\big[\mathbb{W}(P_{W|Z_i}, P_W)\big] \leq \mathbb{E}\big[\mathbb{W}(P_{W|S}, P_W)\big], \tag{8}$$

which is a stronger statement than the original. More precisely, it can be shown that, for all $i \in [n]$,

$$\mathbb{E}\Bigg[\sup_{f\in 1\text{-Lip}(\rho)} \Big\{\mathbb{E}[f(W) \mid Z_i] - \mathbb{E}[f(W)]\Big\}\Bigg] \leq \mathbb{E}\Bigg[\sup_{f\in 1\text{-Lip}(\rho)} \Big\{\mathbb{E}[f(W) \mid S] - \mathbb{E}[f(W)]\Big\}\Bigg],$$

which is equivalent to (8) due to the KR duality.

After writing $\mathbb{E}\big[\sup_{f\in 1\text{-Lip}(\rho)} \big\{\mathbb{E}[f(W) \mid S] - \mathbb{E}[f(W)]\big\}\big]$ in integral form, the result is shown as follows:

$$\begin{aligned}
&\int_{\mathcal{Z}^n} \sup_{f\in 1\text{-Lip}(\rho)} \left\{\int_{\mathcal{W}} f(w) dP_{W|s}(w) - \int_{\mathcal{W}} f(w) dP_W(w)\right\} dP_S(s) \\
&\quad \overset{(a)}{\geq} \int_{\mathcal{Z}} \sup_{f\in 1\text{-Lip}(\rho)} \left\{\int_{\mathcal{Z}^{n-1}} \left(\int_{\mathcal{W}} f(w) dP_{W|s}(w) - \int_{\mathcal{W}} f(w) dP_W(w)\right) P_Z^{\otimes n-1}(s^{-i})\right\} dP_Z(z_i) \\
&\quad \overset{(b)}{=} \int_{\mathcal{Z}} \sup_{f\in 1\text{-Lip}(\rho)} \left\{\int_{\mathcal{W}} f(w) dP_{W|z_i}(w) - \int_{\mathcal{W}} f(w) dP_W(w)\right\} dP_Z(z_i),
\end{aligned}$$

where $s^{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$, (a) is due to the fact that $\sup_g \mathbb{E}[g(X)] \leq \mathbb{E}[\sup_g g(X)]$, and (b) follows from Fubini–Tonelli's theorem and the fact that $P_{W|Z_i} = \mathbb{E}[P_{W|S} \mid Z_i]$.

Finally, noting that (b) is the integral form of $\mathbb{E}\big[\sup_{f\in 1\text{-Lip}(\rho)} \big\{\mathbb{E}[f(W) \mid Z_i] - \mathbb{E}[f(W)]\big\}\big]$ concludes the proof. □

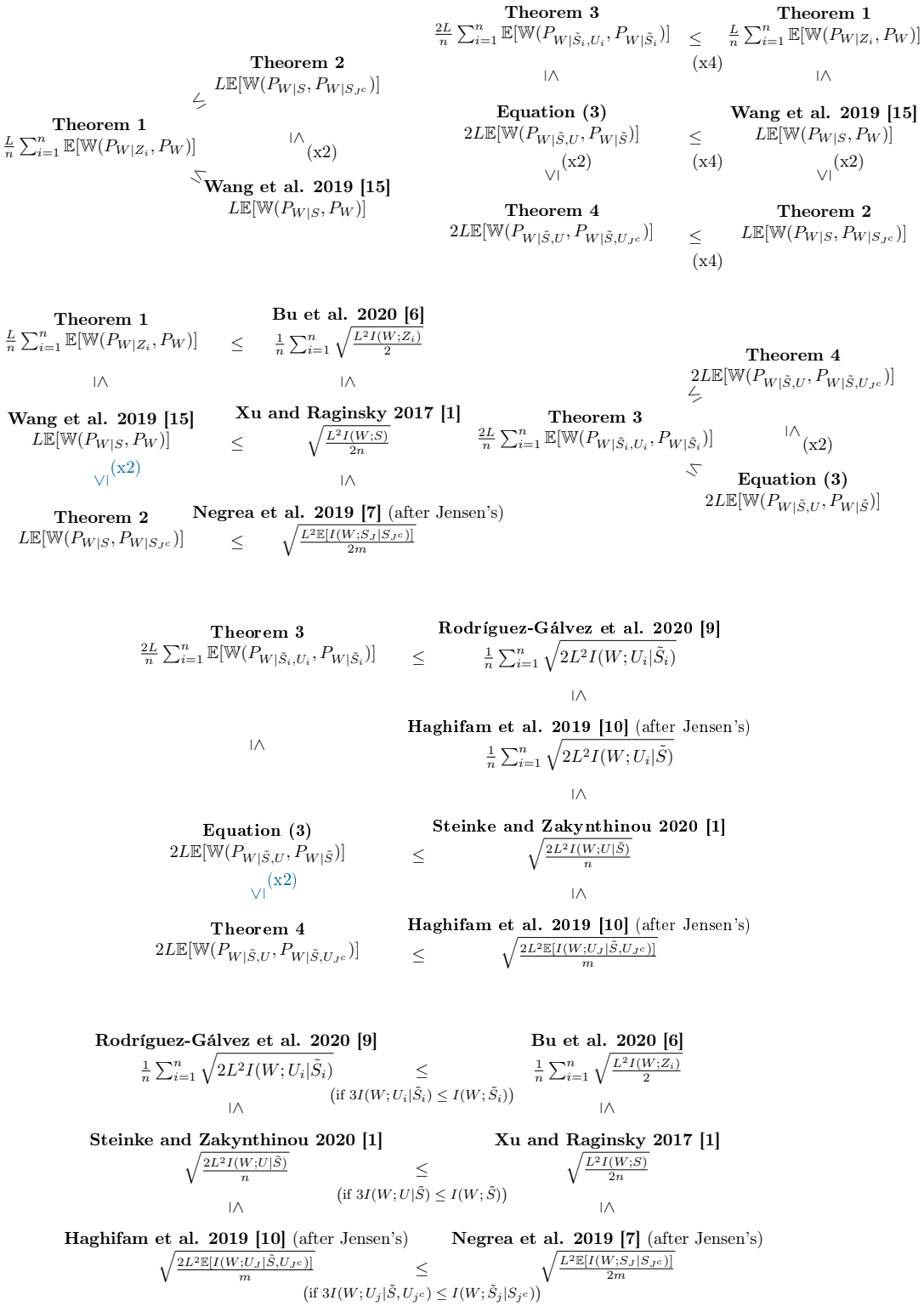

Figure 2: Summary of the comparison between the current and presented bounds based on the Wasserstein distance and the mutual information.

Proposition 3. *Consider the standard setting. Consider also a uniformly random subset of indices $J \subseteq [n]$ of size m. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathbb{W}(P_{W|Z_i}, P_W)\big] \leq \mathbb{E}\big[\mathbb{W}(P_{W|S}, P_{W|S_{J^c}})\big].$$

Proof. The proof of this proposition follows closely that of Proposition 2. First note that the statement may be written as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathbb{W}(P_{W|Z_i}, P_W)\big] \leq \frac{1}{\binom{n}{m}}\sum_{j\in\mathcal{J}}\mathbb{E}\big[\mathbb{W}(P_{W|S}, P_{W|S_{j^c}})\big],$$

where the expectation with respect to P_J has been written explicitly. Then, this result follows by noting that, for all $i \in [n]$ and all $j \subseteq [n]$ such that $i \in j$

$$\mathbb{E}\big[\mathbb{W}(P_{W|Z_i}, P_W)\big] \leq \mathbb{E}\big[\mathbb{W}(P_{W|S}, P_{W|S_{j^c}})\big], \tag{9}$$

which is a stronger statement than the original. This is stronger since one can, without loss of generality, consider the samples Z_i ordered so that the sequence $\{\mathbb{E}[\mathbb{W}(P_{W|Z_i}, P_W)]\}_{i\in[n]}$ is decreasing. Then, $\mathbb{E}[\mathbb{W}(P_{W|Z_1}, P_W)]$ is smaller than $\mathbb{E}[\mathbb{W}(P_{W|S}, P_{W|S_{j^c}})]$ for the $\binom{n-1}{m-1}$ sets j in which sample 1 appears, $\mathbb{E}[\mathbb{W}(P_{W|Z_2}, P_W)]$ is smaller than $\mathbb{E}[\mathbb{W}(P_{W|S}, P_{W|S_{j^c}})]$ for the sets j in which sample 2 appears and sample 1 does not, and so on.

More precisely, it can be shown that, for all $i \in [n]$ and all $j \subseteq [n]$ such that $i \in j$,

$$\mathbb{E}\bigg[\sup_{f\in\text{1-Lip}(\rho)}\Big\{\mathbb{E}[f(W) \mid Z_i] - \mathbb{E}[f(W)]\Big\}\bigg] \leq \mathbb{E}\bigg[\sup_{f\in\text{1-Lip}(\rho)}\Big\{\mathbb{E}[f(W) \mid S] - \mathbb{E}[f(W) \mid S_{j^c}]\Big\}\bigg],$$

which is equivalent to (9) due to the KR duality.

After writing $\mathbb{E}\big[\sup_{f\in\text{1-Lip}(\rho)}\big\{\mathbb{E}[f(W) \mid S] - \mathbb{E}[f(W) \mid S_{j^c}]\big\}\big]$ in integral form, the result is shown as follows:

$$\begin{aligned}
&\int_{\mathcal{Z}^n}\sup_{f\in\text{1-Lip}(\rho)}\bigg\{\int_{\mathcal{W}} f(w)dP_{W|s}(w) - \int_{\mathcal{W}} f(w)dP_{W|s_{j^c}}(w)\bigg\}dP_S(s)\\
&\overset{(a)}{\geq}\int_{\mathcal{Z}}\sup_{f\in\text{1-Lip}(\rho)}\bigg\{\int_{\mathcal{Z}^{n-1}}\bigg(\int_{\mathcal{W}} f(w)dP_{W|s}(w) - \int_{\mathcal{W}} f(w)dP_{W|s_{j^c}}(w)\bigg)P_Z^{\otimes n-1}(s^{-i})\bigg\}dP_Z(z_i)\\
&\overset{(b)}{=}\int_{\mathcal{Z}}\sup_{f\in\text{1-Lip}(\rho)}\bigg\{\int_{\mathcal{W}} f(w)dP_{W|z_i}(w) - \int_{\mathcal{W}} f(w)dP_W(w)\bigg\}dP_Z(z_i),
\end{aligned}$$

where (a) stems from the fact that $\sup_g \mathbb{E}[g(X)] \leq \mathbb{E}[\sup_g g(X)]$, and (b) follows from Fubini–Tonelli's theorem and the fact that $P_{W|Z_i} = \mathbb{E}[P_{W|S} \mid Z_i]$ and $P_W = \mathbb{E}[P_{W|S_{j^c}} \mid Z_i]$, since $i \in j$ and therefore $i \notin j^c$.

Finally, noting that (b) is the integral form of $\mathbb{E}\big[\sup_{f\in\text{1-Lip}(\rho)}\big\{\mathbb{E}[f(W) \mid Z_i] - \mathbb{E}[f(W)]\big\}\big]$ concludes the proof. □

Proposition 4. *Consider the standard setting. Consider also a uniformly random subset of indices $J \subseteq [n]$ of size m. Then,*

$$\mathbb{E}\big[\mathbb{W}(P_{W|S}, P_{W|S_{J^c}})\big] \leq 2\mathbb{E}\big[\mathbb{W}(P_{W|S}, P_W)\big].$$

Proof. An application of the triangle inequality on Wasserstein distances [18, Chapter 6] states that, for all $j \subseteq [n]$ such that $|j| = m$ and all $s \in \mathcal{Z}^n$

$$\mathbb{W}(P_{W|s}, P_{W|s_{j^c}}) \leq \mathbb{W}(P_{W|s}, P_W) + \mathbb{W}(P_{W|s_{j^c}}, P_W). \tag{10}$$

Then, the inequality $\mathbb{E}[\mathbb{W}(P_{W|S_{j^c}}, P_W)] \leq \mathbb{E}[\mathbb{W}(P_{W|S}, P_W)]$ holds by the same arguments of Proposition 2. That is, writing the Wasserstein distance in its KR dual form and noting that the integral of a supremum is greater than the supremum of the integral and that $P_{W|S_{j^c}} = \mathbb{E}[P_{W|S} \mid S_{j^c}]$ for all $j \subseteq [n]$. Hence, taking expectations on both sides of (10) results in

$$\mathbb{E}\big[\mathbb{W}(P_{W|S}, P_{W|S_{J^c}})\big] \leq 2\mathbb{E}\big[\mathbb{W}(P_{W|S}, P_W)\big],$$

which completes the proof. □

D.1.2 Randomized-subsample setting

Proposition 5. *Consider the randomized-subsample setting. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}_i})\big] \leq \mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})\big],$$

where $\tilde{S}_i$ is defined in Theorem 3.

Proof. The proposition follows noting that, for all $i \in [n]$

$$\mathbb{E}[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}})] \leq \mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})], \tag{11}$$

which is a stronger statement than the original. Then, Equation (11) follows by the same arguments as Propositions 2 and 4. That is, writing the Wasserstein distance in its KR dual form and noting that the integral of a supremum is greater than the supremum of the integral and that $P_{W|\tilde{S}_i,U_i} = \mathbb{E}[P_{W|\tilde{S},U} \mid \tilde{S}_i, U_i]$ and $P_{W|\tilde{S}_i} = \mathbb{E}[P_{W|\tilde{S}} \mid \tilde{S}_i, U_i]$. □

Proposition 6. *Consider the randomized-subsample setting. Consider also a uniformly random subset of indices $J \subseteq [n]$ of size m. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}_i})\big] \leq \mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S},U_{J^c}})\big],$$

Proof. Note that the statement of the proposition may be written as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}_i})\big] \leq \frac{1}{\binom{n}{m}}\sum_{j\in\mathcal{J}}\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S},U_{j^c}})\big],$$

where the expectation with respect to P_J has been written explicitly. Then, this result follows by noting that, for all $i \in [n]$ and all $j \subseteq [n]$ such that $i \in j$

$$\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}_i})\big] \leq \mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S},U_{j^c}})\big], \tag{12}$$

which is a stronger statement than the original. Similarly to Proposition 4, this is stronger since one can, without loss of generality, consider the tuple of pairs of samples and their deciding index $(\tilde{S}_i, U_i) = \big((\tilde{Z}_i, \tilde{Z}_{i+n}), U_i\big)$ ordered so that the sequence $\{\mathbb{E}[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}_i})]\}_{i\in[n]}$ is decreasing. Then, $\mathbb{E}[\mathbb{W}(P_{W|\tilde{S}_1,U_1}, P_{W|\tilde{S}_1})]$ is smaller than $\mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S},U_{j^c}})]$ for the $\binom{n-1}{m-1}$ sets j in which the tuple 1 appears, $\mathbb{E}[\mathbb{W}(P_{W|\tilde{S}_2,U_2}, P_{W|\tilde{S}_2})]$ is smaller than $\mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S},U_{j^c}})]$ for the sets j in which tuple 2 appears and tuple 1 does not, and so on.

Then, Equation (12) follows by the same arguments than Propositions 2 and 4. That is, writing the Wasserstein distance in its KR dual form and noting that the integral of a supremum is greater than the supremum of the integral and that $P_{W|\tilde{S}_i,U_i} = \mathbb{E}[P_{W|\tilde{S},U} \mid \tilde{S}_i, U_i]$ and $P_{W|\tilde{S}_i} = \mathbb{E}[P_{W|\tilde{S},U_{j^c}} \mid \tilde{S}_i, U_i]$, since $i \in j$ and therefore $i \notin j^c$. □

Proposition 7. *Consider the randomized-subsample setting. Consider also a uniformly random subset of indices $J \subseteq [n]$ of size m. Then,*

$$\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S},U_{J^c}})\big] \leq 2\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})\big].$$

Proof. An application of the triangle inequality on Wasserstein distances [18, Chapter 6] states that, for all $j \subseteq [n]$ such that $|j| = m$, all $\tilde{s} \in \mathcal{Z}^{2n}$, and all $u \in [0,1]^n$

$$\mathbb{W}(P_{W|\tilde{s},u}, P_{W|\tilde{s},u_{j^c}}) \leq \mathbb{W}(P_{W|\tilde{s},u}, P_{W|\tilde{s}}) + \mathbb{W}(P_{W|\tilde{s},u_{j^c}}, P_{W|\tilde{s}}). \tag{13}$$

Then, the inequality $\mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U_{j^c}}, P_{W|\tilde{S}})] \leq \mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})]$ holds by the same arguments of Proposition 2. That is, writing the Wasserstein distance in its KR dual form and noting that the integral of a supremum is greater than the supremum of the integral and that $P_{W|\tilde{S},U_{j^c}} = \mathbb{E}[P_{W|\tilde{S},U} \mid \tilde{S}, U_{j^c}]$ for all $j \subseteq [n]$. Hence, taking expectations on both sides of (13) results in

$$\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S},U_{J^c}})\big] \leq 2\mathbb{E}\big[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})\big],$$

which completes the proof. □

D.1.3 Comparison between the settings

Proposition 8. *Consider the standard and the randomized-subsample settings. Consider also a uniformly random subset of indices $J \subseteq [n]$ of size m. Then,*

$$\mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})] \leq 2\mathbb{E}[\mathbb{W}(P_{W|S}, P_W)],$$
$$\mathbb{E}[\mathbb{W}(P_{W|\tilde{S}_i,U_i}, P_{W|\tilde{S}_i})] \leq 2\mathbb{E}[\mathbb{W}(P_{W|Z_i}, P_W)], \text{ and}$$
$$\mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S},U_{J^c}})] \leq 2\mathbb{E}[\mathbb{W}(P_{W|S}, P_{W|S_{J^c}})].$$

Proof. The proofs of the three statements are analogous. Therefore only the proof of the first statement is explicitly written.

An application of the triangle inequality on Wasserstein distances [18, Chapter 6] states that, for $\tilde{s} \in \mathcal{Z}^{2n}$ and all $u \in [0,1]^n$ such that s is obtained through $\tilde{s}$ and u as explained in the introduction,

$$\mathbb{W}(P_{W|\tilde{s},u}, P_{W|\tilde{s}}) \leq \mathbb{W}(P_{W|\tilde{s},u}, P_W) + \mathbb{W}(P_{W|\tilde{s}}, P_W). \tag{14}$$

Then, the inequality $\mathbb{E}[\mathbb{W}(P_{W|\tilde{S}}, P_W)] \leq \mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_W)]$ holds by the same arguments of Proposition 2. That is, writing the Wasserstein distance in its KR dual form and noting that the integral of a supremum is greater than the supremum of the integral and that $P_{W|\tilde{S}} = \mathbb{E}[P_{W|\tilde{S},U} \mid \tilde{S}]$. Hence, taking expectations on both sides of (14) and noting that $P_{\tilde{S},U} = P_S$ almost surely results in

$$\mathbb{E}[\mathbb{W}(P_{W|\tilde{S},U}, P_{W|\tilde{S}})] \leq 2\mathbb{E}[\mathbb{W}(P_{W|S}, P_W)],$$

which completes the proof. □

D.2 Comparison of the mutual information based bounds

D.2.1 Standard setting

In the standard setting, after a further application of Jensen's inequality, the bounds derived from Corollary 1 or [6, Proposition 1] are the tightest, followed by [4, Theorem 1] and the bounds derived from Corollary 2 when $|J| = 1$ or [7, Theorem 2.4] if the arbitrary random variable R is not considered. This follows since by [6, Proposition 2] and [38, Lemma 3.7] or [9, Lemma 2]

$$\sum_{i=1}^{n} I(W;Z_i) \leq I(W;S) \leq \sum_{i=1}^{n} I(W;Z_i|S^{-i}),$$

where $S^{-i} = S \setminus Z_i$. More generally, with trivial modifications to the proof from [9, Lemma 2], it can be shown that

$$\sum_{i=1}^{n} I(W;Z_i) \leq I(W;S) \leq \mathbb{E}[I(W;S_J|S_{J^c})],$$

noting that, for any $j \subseteq [n]$ of size m, if the elements of s_j are ordered as $Z_{k_1}, \ldots, Z_{k_m}$,

$$I(W;S) = \sum_{i=1}^{n} I(W;Z_i|S^{i-1}) \leq \frac{1}{\binom{n}{m}} \sum_{\iota=1}^{m} I(W;Z_{k_\iota}|S_{j^c}, S_j^{k_{\iota-1}}),$$

since every time that $i = k_\iota$ then $S^{i-1} \subseteq (S_{j^c} \cup S_j^{k_{\iota-1}})$, where $S_j^{k_{\iota-1}}$ are the first ι-1 elements of S_j.

D.2.2 Randomized-subsample setting

With a similar argument to the one for the standard setting, after a further application of Jensen's inequality, the bounds derived from [10, Theorem 3.4] are tighter than [1, Theorem 5.1], which in turn are tighter than the bounds derived from Corollary 4 when $|J| = 1$ or [10, Theorem 3.7] if the arbitrary random variable R is not considered, since

$$\sum_{i=1}^{n} I(W;U_i|\tilde{S}) \leq I(W;U|\tilde{S}) \leq \mathbb{E}[I(W;U_J|\tilde{S}, U_{J^c})].$$

Furthermore, the bounds derived from Corollary 3 or [9, Proposition 3] are the tightest as dictated by [9, Lemma 3] or [12, Lemma 2]. This way, the relationship between the conditional mutual information terms is

$$\sum_{i=1}^{n} I(W;U_i|\tilde{Z}_i,\tilde{Z}_{i+n}) \leq \sum_{i=1}^{n} I(W;U_i|\tilde{S}) \leq I(W;U|\tilde{S}) \leq \mathbb{E}[I(W;U_J|\tilde{S},U_{J^c})].$$

D.2.3 Comparison between the settings

Similarly, one may note that $I(W;U|\tilde{S}) \leq I(W;S)$ by [10], $I(W;U_i|\tilde{Z}_i,\tilde{Z}_{i+n}) \leq I(W;Z_i)$, and $I(W;U_j|\tilde{S}) \leq I(W;U_j|\tilde{S},U_{j^c}) \leq I(W;S_j|S_{j^c})$ for any subset of indices $j \subseteq [n]$. Nonetheless, the additional factor of two in the bounds of the randomized-subsample setting makes the comparison between the bounds of the different settings harder.

An attempt for this comparison is given in [8], where they note that, since $(U,\tilde{S}) \leftrightarrow S \leftrightarrow W$ form a Markov chain and S is a deterministic function of $\tilde{S}$ and U, then $I(W;S) = I(W;\tilde{S}) + I(W;U|\tilde{S})$, and hence the bound from the randomized-subsample setting [1, Theorem 5.1] is tighter than the one from the standard setting [4, Theorem 1] if $3I(W;U|\tilde{S}) \leq I(W;\tilde{S})$. There are similar requirements for the single-letter and random-subset bounds, namely:

- The bound derived from Corollary 3 or [9, Proposition 3] is tighter than [6, Proposition 1] if $3I(W;U_i|\tilde{Z}_i,\tilde{Z}_{i+n}) \leq I(W;\tilde{Z}_i,\tilde{Z}_{i+n})$.
- The bound derived from Corollary 4 or [10, Theorem 3.7] is tighter than [7, Theorem 2.4] if $3I(W;U_j|\tilde{S},U_{j^c}) \leq I(W;\tilde{S}_j|S_{j^c})$.

Remark 4. *Note that, sometimes, seemingly looser bounds can lead to tighter or more tractable bounds for specific algorithms. For instance, the random-subset bounds from [7, 9, 10] lead to tighter Langevin dynamics and stochastic gradient Langevin dynamics than the single-letter bounds from [6].*

E Derivations for the Gaussian location model example

The problem considered in the example is the estimation of the mean μ of a d-dimensional Gaussian distribution with known covariance matrix $\sigma^2 I_d$. Furthermore, there are n samples $S = (Z_1,\dots,Z_n)$ available, the loss is measured with the Euclidean distance $\ell(w,z) = \|w-z\|_2$, and the estimation is their empirical mean $W = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

To calculate the expected generalization error and derive different bounds, it is convenient to know how the random variables are distributed. For example, in this setting $P_Z = \mathcal{N}(\mu,\sigma^2 I_d)$, $P_W = \mathcal{N}\Big(\mu,\frac{\sigma^2}{n}I_d\Big)$, $P_{W|Z_i} = \mathcal{N}\Big(\frac{(n-1)\mu+Z_i}{n},\frac{\sigma^2(n+1)}{n^2}I_d\Big)$, $P_{W|S^{-j}} = \mathcal{N}\Big(\frac{\mu}{n}+\frac{1}{n}\sum_{i\neq j}Z_i,\sigma^2 I_d\Big)$, and $P_{W|S} = \delta\Big(\frac{1}{n}\sum_{i=1}^{n}Z_i\Big)$. Another important feature of this problem is that the loss function is 1-Lipschitz under $\rho(w,w') = \|w-w'\|_2$.

E.1 Expected generalization error

In order to derive an exact expression of the generalization error, it is suitable to write it in the following explicit form:

$$\overline{\text{gen}}(W,S) = \mathbb{E}[\ell(W,Z)] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\ell(W,Z_i)],$$

where $Z \sim P_Z$ is independent of W. Then, the two terms can be evaluated independently.

The first term is equivalent to

$$\mathbb{E}[\ell(W,Z)] = \mathbb{E}[\|W-Z\|_2] = \sqrt{2\sigma^2\Big(1+\frac{1}{n}\Big)}\frac{\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)},$$

where the first equality follows from the definition of the loss function. The second equality follows from noting that $(W - Z) \sim \mathcal{N}\big(0, \sigma^2\big(1 + \frac{1}{n}\big) I_d\big)$ and therefore $\|W - Z\|_2 = \sqrt{\sigma^2\big(1 + \frac{1}{n}\big)} X$, where X is distributed according to the chi distribution with d degrees of freedom.

Similarly, the summands of the second term are equivalent to

$$\mathbb{E}[\ell(W, Z_i)] = \mathbb{E}[\|W - Z_i\|_2] = \sqrt{2\sigma^2\Big(1 - \frac{1}{n}\Big)} \frac{\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)},$$

where as before the first equality follows from the definition of the loss function. The second equality follows from noting that $(W - Z_i) \sim \mathcal{N}\big(0, \sigma^2\big(1 - \frac{1}{n}\big) I_d\big)$ and therefore $\|W - Z_i\|_2 = \sqrt{\sigma^2\big(1 - \frac{1}{n}\big)} X$, where X is distributed according to the chi distribution with d degrees of freedom. In this case, W and Z_i are not independent random variables. In fact, (W, Z_i) is normally distributed with covariance matrix

$$\begin{pmatrix} \frac{\sigma^2}{n} I_d & \frac{\sigma^2}{n} I_d \\ \frac{\sigma^2}{n} I_d & \sigma^2 I_d \end{pmatrix},$$

from which the distribution of $W - Z_i$ is deduced.

Finally, subtracting both terms results in

$$\overline{\text{gen}}(W, S) = \sqrt{\frac{2\sigma^2}{n}} \Big(\sqrt{n+1} - \sqrt{n-1}\Big) \frac{\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)} \leq \frac{\sqrt{2\sigma^2 d}}{n}.$$

where the inequality follows from the following two bounds: (i) $\sqrt{n+1} - \sqrt{n-1} \leq \sqrt{\frac{2}{n}}$, which is obtained by multiplying and dividing by $\sqrt{n+1} + \sqrt{n-1}$ and noting that $\sqrt{n+1} + \sqrt{n-1} \geq \sqrt{2n}$, and (ii) the upper bound on the ratio of gamma distributions by $\sqrt{\frac{d}{2}}$ using the series expansion at $d \to \infty$.

E.2 Wasserstein distance bound

The bound from [15] can be calculated exactly since $P_{W|S}$ is a delta distribution, that is

$$\mathbb{E}\big[\mathbb{W}(P_{W|S}, P_W)\big] = \mathbb{E}\left[\Big\| \frac{1}{n} \sum_{i=1}^{n} Z_i - \frac{1}{n} \sum_{i=1}^{n} Z_i' \Big\|_2\right] = \sqrt{\frac{4\sigma^2}{n}} \frac{\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)} \leq \sqrt{\frac{2\sigma^2 d}{n}}$$

where $Z_i' \sim P_Z$ are independent copies of Z_i. Hence, the difference is distributed as a normal distribution with mean 0 and covariance $\frac{2\sigma^2}{n} I_d$, which means that the norm is $\sqrt{\frac{2\sigma^2}{n}} X$, where X is a chi random variable with d degrees of freedom.

E.3 Individual sample Wasserstein distance bound

An exact calculation of the bound from Theorem 1 is cumbersome. However, the Wasserstein distance of order one can be bounded from above by the Wasserstein distance of order two (Remark 1), which has a closed form expression for Gaussian distributions. More specifically,

$$\mathbb{E}\big[\mathbb{W}(P_{W|Z_i}, P_W)\big] \leq \mathbb{E}\big[\mathbb{W}_2(P_{W|Z_i}, P_W)\big] \leq \frac{\sqrt{2\sigma^2}}{n} \frac{\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)} + \sqrt{\frac{\sigma^2 d}{n^3}} \leq \frac{\sqrt{\sigma^2 d}}{n} + \sqrt{\frac{\sigma^2 d}{n^3}}.$$

The second inequality follows from the closed-form expression for the squared Wasserstein distance of order 2, namely

$$\mathbb{W}(P_{W|Z_i}, P_W)^2 = \frac{1}{n^2} \|\mu - Z_i\|^2 + \frac{\sigma^2 d}{n} \Big(1 + \frac{n-1}{n} - 2\sqrt{\frac{n-1}{n}}\Big),$$

where the term $(1 + \frac{n+1}{n} + \sqrt{\frac{n+1}{n}})$ is a perfect square that is bounded from above by $\frac{1}{n^2}$. Then the expression results from employing the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and noting that $\|\mu - Z_i\|_2 = \sigma X$, where X is a chi distributed random variable with d degrees of freedom.

E.4 Random subset Wasserstein distance bound

As in E.2, since $P_{W|S}$ is a delta distribution, the bound from Theorem 2 can be calculated exactly. In particular, the bound assuming that $|J| = 1$ is

$$\mathbb{E}\big[\mathbb{W}(P_{W|S}, P_{W|S^{-J}})\big] = \mathbb{E}\left[\Big\|\frac{1}{n}\sum_{i=1}^{n} Z_i - \Big(\frac{Z'_J}{n} + \frac{1}{n}\sum_{i\neq J} Z_i\Big)\Big\|_2\right] = \frac{\sqrt{4\sigma^2}}{n}\frac{\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)} \leq \frac{\sqrt{2\sigma^2 d}}{n},$$

where $Z'_J \sim P_Z$ is an independent copy of Z_J. Hence the norm is $\frac{\sqrt{2\sigma^2}}{n}X$, where X is a chi random variable with d degrees of freedom.

E.5 Individual sample mutual information bound

The individual sample mutual information is $I(W; Z_i) = \frac{d}{2}\log\big(\frac{n}{n-1}\big)$ for all $i \in [n]$ [6]. Nonetheless, in order to employ the bound from [6], the loss function $\ell(W, Z)$ needs to have a cumulant generating function $\Lambda(\lambda)$ bounded from above by a convex function $\psi(\lambda)$ such that $\psi(0) = \psi'(0) = 0$ for all $\lambda \in (-b, 0]$ for some $b \in \mathbb{R}_+$, where $Z \sim P_Z$ is independent of W.

The loss function $\ell(W, Z) = \|W - Z\|_2$ is $\sqrt{\sigma^2(1 + \frac{1}{n})}X$, where X is a chi random variable with d degrees of freedom. The moment generating function $M(\lambda)$ of such a random variable is

$$M(\lambda) = \bar{M}\Big(\frac{d}{2}, \frac{1}{2}, \frac{\lambda^2}{2}\Big) + \frac{\lambda\sqrt{2}\Gamma\big(\frac{d+1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)}\bar{M}\Big(\frac{k+1}{2}, \frac{3}{2}, \frac{\lambda^2}{2}\Big),$$

where $\bar{M}$ is the Kummer's confluent hypergeometric function.

The expression of this moment generating function is too convoluted to study for $d > 1$. Nonetheless, for $d = 1$ it has a closed form expression, namely

$$M(\lambda) = e^{\frac{\lambda^2}{2}}\Big(1 + \operatorname{erf}\Big(\frac{\lambda}{\sqrt{2}}\Big)\Big).$$

Therefore, the cumulant generating function is $\Lambda(\lambda) = \frac{\lambda^2}{2} + \log(1 + \operatorname{erf}(\frac{\lambda}{\sqrt{2}}))$, which is bounded from above by the convex function $\psi(\lambda) = \frac{\lambda^2}{2}$ for all $\lambda \in (-\infty, 0]$. Hence, the bound from [6] can be applied yielding

$$\overline{\text{gen}}(W, S) \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2\Big(1 + \frac{1}{n}\Big)I(W; Z_i)} \leq \sqrt{\sigma^2\Big(1 + \frac{1}{n}\Big)\log\Big(\frac{n}{n-1}\Big)} \leq \sqrt{\frac{2\sigma^2}{n-1}},$$

where the last inequality stems from noting that $\frac{n}{n-1} = 1 + \frac{1}{n-1}$, the fact that $\log(1 + x) \leq x$, and bounding $(1 + \frac{1}{n})$ from above by 2.

F Randomized-subsample setting and the BH inequality

In Corollaries 3 and 4, the immediate bound that stems from the use of the BH inequality is not included. The reason for this is that the relative entropies $D_{\mathrm{KL}}(P_{W|\tilde{Z}_i,\tilde{Z}_{i+n},U_i} \,\|\, P_{W|\tilde{Z}_i,\tilde{Z}_{i+n}})$ and $D_{\mathrm{KL}}(P_{W|\tilde{S},U,R} \,\|\, P_{W|\tilde{S},U_{J^c},R})$, when $|J| = 1$, are never greater than $\log(2)$ as shown in Lemma 3 below. Hence, the range of these relative entropies is inside the range where Pinsker's inequality is tighter than the BH inequality.

Lemma 3. *Let $P_{X|A,B}$ be a conditional probability distribution on $\mathcal{X}$, where B is a Bernoulli random variable with probability $1/2$ and $A \in \mathcal{A}$ is a random variable independent of B. Let also $P_{X|A} = \mathbb{E}[P_{X|A,B} \mid A]$. Then, $D_{\mathrm{KL}}(P_{X|A,B} \,\|\, P_{X|A}) \leq \log(2)$.*

Proof. In this situation $P_{X|A}$ dominates $P_{X|A,B}$ and $P_{X|A,(1-B)}$, that is $P_{X|A,B} \ll P_{X|A}$ and $P_{X|A,(1-B)} \ll P_{X|A}$, since $P_{X|A} = (P_{X|A,B} + P_{X|A,(1-B)})/2$. Therefore,

$$\begin{aligned} D_{\mathrm{KL}}(P_{X|A,B} \,\|\, P_{X|A}) &= \mathbb{E}\left[\log\left(\frac{dP_{X|A,B}}{d\left(\frac{1}{2}P_{X|A,B} + \frac{1}{2}P_{X|A,(1-B)}\right)} \right) \,\middle|\, A, B \right] \\ &\overset{(a)}{=} -\mathbb{E}\left[\log\left(\frac{d\left(\frac{1}{2}P_{X|A,B} + \frac{1}{2}P_{X|A,(1-B)}\right)}{dP_{X|A,B}} \right) \,\middle|\, A, B \right] \\ &\overset{(b)}{=} \log(2) - \mathbb{E}\left[\log\left(1 + \frac{dP_{X|A,(1-B)}}{dP_{X|A,B}} \right) \,\middle|\, A, B \right] \\ &\overset{(c)}{\leq} \log(2), \end{aligned}$$

where (a) stems from [39, Exercise 9.27], (b) follows from the linearity $P_{X|A,B}$-a.e. of the Radon–Nikodym derivative, and (c) is due to the fact that $\log(1+x) \geq 0$ for all $x \geq 0$ and the fact that $dP_{X|A,(1-B)}/dP_{X|A,B}$ is always positive. Also, steps (a), (b), and (c) are possible since the expectation integrates over the support of $P_{X|A,B}$, avoiding the problems of absolute continuity in (c) and absorbing the $P_{X|A,B}$-a.e. properties for (a) and (b). □

Remark 5. *Lemma 3 can be easily extended to the case where B is a sequence of k Bernoulli random variables B_i, noting that $P_{X|A} = 2^{-k}\sum_{j=1}^{2^k} P_{X|A,\mathcal{B}_j}$, where $\mathcal{B}$ are all the 2^k random sequences $\mathcal{B}_j$ where the i-th element can be either B_i or $(1-B_i)$. In that case, we have that $D_{\mathrm{KL}}(P_{X|A,B} \,\|\, P_{X|A}) \leq k\log(2)$.*

Then, note that $D_{\mathrm{KL}}(P_{W|\tilde{Z}_i,\tilde{Z}_{i+n},U_i} \,\|\, P_{W|\tilde{Z}_i,\tilde{Z}_{i+n}}) \leq \log(2)$ if $A = (\tilde{Z}_i, \tilde{Z}_{i+n})$, $B = U_i$, and $X = W$. Similarly, for $|J| = 1$, note that $D_{\mathrm{KL}}(P_{W|\tilde{S},U,R} \,\|\, P_{W|\tilde{S},U_{J^c},R}) \leq \log(2)$ if $A = (\tilde{S}, U_{J^c}, R)$, $B = U_J$, and $X = W$.

Remark 6. *In Corollary 4, when $|J| > 2$, it is not guaranteed that the inequality obtained from Pinsker's inequality is tighter than the one obtained with the BH inequality. For instance, as per Remark 5, $D_{\mathrm{KL}}(P_{W|\tilde{S},U,R} \,\|\, P_{W|\tilde{S},U_{J^c},R})$ could be as large as $|J|\log(2)$, which is already larger than 1.6 for $|J| = 3$. Hence, for $|J| > 2$, one should also consider that inequality if one desires the tightest bound.*

However, the bound derived from the BH inequality was not included since this kind of bounds are usually employed for $|J| = 1$, e.g., [10, Theorem 4.2] and [9, Proposition 6]. Moreover, after applying Jensen's inequality, it is shown that the derived mutual-information bounds are the tightest when $|J| = 1$ [10, Corollary 3.3].

G Rate-distortion theory and generalization

G.1 Rate–distortion theory

Rate–distortion theory [22, 19] deals with the problem of determining the minimum number of bits, determined by the *rate* R, that should be employed to characterize a signal X by Y so that this signal can later be recovered with an expected *distortion* lower than δ. Formally, given a signal X with distribution P_X and a distortion measure d, the *rate–distortion* function $R(\delta)$ finds the optimal encoding distribution $P^{\star}_{Y|X}$, i.e., the channel $P_{Y|X}$ that generates a representation Y with the minimum amount of bits $R(\delta)$ and an expected distortion lower than δ. Namely,

$$R(\delta) = \inf_{P_{Y|X}:\mathbb{E}[d(X,Y)]\leq\delta} I(X;Y).$$

A celebrated result in rate–distortion theory is its duality. More precisely, instead of looking for the channel $P_{Y|X}$ that most compresses a signal X with a limited distortion δ, one can look for the channel $P_{Y|X}$ that less distorts the signal X with a limited budget of bits r. That is, one can solve the *distortion–rate* function

$$D(r) = \inf_{P_{Y|X}:I(X;Y)\leq r} \mathbb{E}[d(X,Y)].$$

Then, the duality theorem states that $R(\delta) = D^{-1}(\delta)$ and $D(r) = R^{-1}(r)$ [40, Lemma 4.1.2], or, in words, that the inverse of the rate–distortion function is the distortion–rate function, and vice versa.

Remark 7. *Rate–distortion theory is a well-studied field and the rate–distortion and distortion–rate functions have many more interesting properties and analytical solutions and bounds for particular (and common) cases.*

Single-letterization Sometimes, if X is a sequence of signals $X = (X_1, \dots, X_n)$, solving the rate–distortion function is challenging. Assume that the signals X_i are independent and the distortion d is separable, i.e., $\mathbb{E}[d(X, Y)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[d(X_i, Y_i)]$. Then, a simpler task is to solve the single-letter version of the rate–distortion function. Namely, for any $i \in [n]$, to solve

$$R_i(\delta_i) = \inf_{P_{Y_i|X_i}:\mathbb{E}[d(X_i,Y_i)]\leq\delta} I(X_i; Y_i).$$

Then, for any i, the equality $nR_i(\delta) = R(\delta)$ [19, Theorem 26.1] holds. Hence, the channel $P_{Y|X} = P_{Y_i|X_i}^{\otimes n}$ can be used as a proxy for $P_{Y|X}^{\star}$.

Backward-channel Sometimes, working with the backward channel $P_{X|Y}$ is convenient to derive an analytical solution to the rate–distortion function for a particular input signal distribution, e.g., Bernoulli or Gaussian [19, Chapter 27.1]; to derive an analytical bound for certain distribution families, e.g., bounded variance; or to derive an analytical bound for certain distortion families, e.g., difference, additive, or autoregressive distortions [40, Sections 4.3, 4.6, and 4.7]

G.2 Connection to the generalization error

When designing a learning algorithm $P_{W|S}$, the aim is an algorithm that attains a low accuracy error (or risk) while generalizing well. This informal sentiment can be posed as constrained optimization as follows: to design an algorithm $P_{W|S}$ that has the minimum possible expected generalization error $G(\epsilon)$ while maintaining a population risk lower than ϵ. Namely, one may consider the *generalization–risk* function to be

$$G(\epsilon) = \inf_{P_{W|S}:\mathbb{E}[L_S(W)]\leq\epsilon} \mathbb{E}[\mathrm{gen}(W, S)],$$

and select the algorithm $P_{W|S}^{\star}$ that solves it.

Furthermore, the expected generalization error increases with $I(W; S)$ as shown in [4, Theorem 1] or (2). Therefore, one may instead consider the *information-generalization–risk* function

$$G^{\blacktriangle}(\epsilon) = \inf_{P_{W|S}:\mathbb{E}_{P_{W,S}}[L_S(W)]\leq\epsilon} I(W; S),$$

and use it as a surrogate of the generalization–risk function to choose the algorithm $P_{W|S}^{\blacktriangle}$ that solves $G^{\blacktriangle}(\epsilon)$ as a proxy for $P_{W|S}^{\star}$. Therefore, the powerful rate–distortion theory may be employed to select a sensible learning algorithm or, at least, to better understand the trade-off between generalization and risk.

Remark 8. *There are several issues to be considered when applying rate–distortion theory to the information-generalization–risk function. For instance, usually a hypothesis is not separable, i.e., $W \neq (W_1, \dots, W_n)$, and hence the single-letterization of the problem must be carried obtaining $P_{W|Z_i}$ instead of $P_{W_i|Z_i}$, which is inconvenient since then obtaining $P_{W|S}$ from $P_{W|Z_i}$ is cumbersome. Nonetheless, the aim of this section is just to give some intuition about the connection between rate–distortion and generalization, and a proper formal framework is beyond the objective of this paper. The reader is referred to [11] and [23] for other connections between these two concepts.*

H Additional remarks on the chi-squared based bounds

As mentioned in §3.4, the presented bounds in terms of the total variation result in bounds based on the χ^2-divergence and other f-divergences employing the joint range strategy [19, Chapter 7]. In order to do so, the loss function ℓ is required to be L-Lipschitz for all $z \in \mathcal{Z}$ under the discrete metric, or, in other words, to be bounded in a range $[c, c + L]$ for some $c \in \mathbb{R}$.

Claim 1. *If a function f is L-Lipschitz under the discrete metric, it is bounded in $[c, c+L]$ for some $c \in \mathbb{R}$.*

Proof. If $|f(x) - f(y)| \leq L\rho_{\mathrm{H}}(x, y)$ for all $x, y \in \mathcal{X}$ then $|f(x) - f(y)| \leq L$. This holds if and only if $f : \mathcal{X} \to [c, c+L]$ for some $c \in \mathbb{R}$. □

As an example, Equation (6) is obtained as a corollary of Theorem 1. However, note that Theorems 1, 2, 3, and 4, and Equations (1) and (3) can be replicated using the variational representation of the χ^2-divergence [41, Example 6.4], which states that for all distributions P, Q over $\mathcal{X}$

$$\chi^2(P, Q) = \sup_{f:\mathcal{X}\to\mathbb{R}} \left\{ \frac{\big(\mathbb{E}[f(X)] - \mathbb{E}[f(X)]\big)^2}{\mathrm{Var}[Y]} \right\},$$

where $X \sim P$, $Y \sim Q$, and the supremum is taken over all functions f with finite expectation with respect to P and Q and finite variance with respect to Q. Using this tool instead of the KR duality, Theorem 1 would result in

$$\big|\overline{\mathrm{gen}}(W, S)\big| \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\sqrt{\mathrm{Var}\big[\ell(W', Z_i)\big]\, \chi^2(P_{W|Z_i}, P_W)} \right], \tag{15}$$

where W' is an independent copy of W such that $P_{W,Z_i} = P_W \otimes P_{Z_i}$ for all $i \in [n]$. Equation (15), instead of requiring that the loss ℓ is L-Lipschitz under the discrete metric for all $z \in \mathcal{Z}$, it requires that the function $\ell(w, z)$ has finite expectation with respect to $P_{W|Z_i=z}$ and P_W and finite variance with respect to P_W for all $z \in \mathcal{Z}$. Note that this is a weaker requirement since Lipschitzness under the discrete metric implies boundedness (Claim 1) and Popoviciu's inequality states that a bounded random variable has finite variance.

In particular, under the assumption that ℓ is bounded with a range $[c, L+c]$, one may use Popoviciu's inequality [42], i.e., $\mathrm{Var}[\ell(W', Z_i)] \leq L^2/4$, to recover the corollaries of Theorems 1, 2, 3, and 4 using the discrete metric and the joint range. As an example, using Popoviciu's inequality in combination with (15) recovers (6). Hence, the bounds obtained through the variational representation of the χ^2-divergence are both tighter and more general than those obtained after applying the joint range strategy as in [43, §IV-B] to the total variation bounds obtained through the KR duality. Compared to the bound from (5), obtained after applying first Pinsker's inequality and then the joint range strategy, Equation (15) becomes looser as soon as

$$\chi^2(P_{W|Z_i}, P_W) \geq \frac{-1}{\alpha}\Big(\mathtt{W}\big(-\alpha e^{-\alpha}\big) + \alpha\Big),$$

where $\mathtt{W}$ denotes the Lambert or product-log function and $\alpha = \mathrm{Var}[\ell(W', Z_i)]/2$. In the extreme case where $\mathrm{Var}[\ell(W', Z_i)] = L^2/4$, Equation (5) is tighter than (15) as soon as $\chi^2(P_{W|Z_i}, P_W) \gtrapprox 2.51$, a restriction that becomes more favorable to (15) as the variance decreases. More precisely, the bound obtained from the variational representation of the χ^2-divergence is tighter than (5) in the range where both bounds are non-vacuous, i.e., when $\mathrm{Var}[\ell(W', Z_i)] \leq (e^2 - 1)^{-1} L^2$.