Generalization Bounds via Information Density and Conditional Information Density

Citation for the original published paper (version of record):
Hellström, F., Durisi, G. (2020). Generalization Bounds via Information Density and Conditional Information Density. IEEE Journal on Selected Areas in Information Theory, 1(3): 824-839.
http://dx.doi.org/10.1109/JSAIT.2020.3040992

Generalization Bounds via Information Density and Conditional Information Density

Fredrik Hellström, *Student Member, IEEE*, Giuseppe Durisi, *Senior Member, IEEE* .

***Abstract*—We present a general approach, based on an exponential inequality, to derive bounds on the generalization error of randomized learning algorithms. Using this approach, we provide bounds on the average generalization error as well as bounds on its tail probability, for both the PAC-Bayesian and single-draw scenarios. Specifically, for the case of sub-Gaussian loss functions, we obtain novel bounds that depend on the information density between the training data and the output hypothesis. When suitably weakened, these bounds recover many of the information-theoretic bounds available in the literature. We also extend the proposed exponential-inequality approach to the setting recently introduced by Steinke and Zakynthinou (2020), where the learning algorithm depends on a randomly selected subset of the available training data. For this setup, we present bounds for bounded loss functions in terms of the conditional information density between the output hypothesis and the random variable determining the subset choice, given all training data. Through our approach, we recover the average generalization bound presented by Steinke and Zakynthinou (2020) and extend it to the PAC-Bayesian and single-draw scenarios. For the single-draw scenario, we also obtain novel bounds in terms of the conditional α-mutual information and the conditional maximal leakage.**

I. Introduction

A randomized learning algorithm $P_{W|\boldsymbol{Z}}$ consists of a probabilistic mapping from a set of training data $\boldsymbol{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$, which we assume to have been generated independently from an unknown distribution P_Z on the instance space $\mathcal{Z}$, to an output hypothesis $W \in \mathcal{W}$, where $\mathcal{W}$ is the hypothesis space. The goal is to find a hypothesis W that results in a small expected loss $L_{P_Z}(W) = \mathbb{E}_{P_Z}[\ell(W, Z)]$, where $\ell(\cdot, \cdot)$ is some suitably chosen loss function. A typical strategy to achieve this goal is *empirical risk minimization*, according to which W is selected so as to minimize the empirical loss $L_{\boldsymbol{Z}}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i)$. A central objective in statistical learning theory is to determine when this choice results in a small population loss $L_{P_Z}(W)$. To this end, one seeks to bound the *generalization error*, defined as $\mathrm{gen}(W, \boldsymbol{Z}) = L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)$. Since the learning algorithm is randomized, bounds on $\mathrm{gen}(W, \boldsymbol{Z})$ can come in several flavors. One possibility is to bound the average generalization error $|\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})]|$. In practice, one might be more interested in an upper bound on $\left|\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})]\right|$ that holds with probability at least $1 - \delta$ under the product distribution $P_{\boldsymbol{Z}}$. Here, $\delta \in (0, 1)$ is the so-called confidence parameter. Bounds of this type, which are typically referred to as *probably approximately correct (PAC)-Bayesian* bounds [2], [3], are relevant for the scenario in which a new hypothesis W is drawn from $P_{W|\boldsymbol{Z}}$ every time the algorithm is used. For the scenario in which W is drawn from $P_{W|\boldsymbol{Z}}$ only once—a setup that, following the terminology in [4], we shall refer to as *single-draw*—one may instead be interested in obtaining an upper bound on $|\mathrm{gen}(W, \boldsymbol{Z})|$ that holds with probability at least $1 - \delta$ under the joint distribution $P_{W\boldsymbol{Z}}$. If the dependence of a probabilistic bound (PAC-Bayesian or single-draw) on δ^{-1} is at most logarithmic, the bound is usually referred to as a *high-probability* bound. Furthermore, a probabilistic bound is termed *data-independent* if it does not depend on the specific instance of $\boldsymbol{Z}$, and *data-dependent* if it does. Data-independent bounds allow one to characterize the sample complexity [5, p. 44], defined as the minimum number of training samples needed to guarantee that the generalization error is within a desired range, with a desired confidence level. However, data-dependent results are often tighter. Indeed, many of the available data-independent bounds can be recovered as relaxed versions of data-dependent bounds.

Classical PAC bounds on the generalization error, such as those based on the Vapnik-Chervonenkis (VC) dimension [5, p. 67], are probabilistic bounds of a stronger variety than the PAC-Bayesian and single-draw bounds just introduced. Indeed, they hold uniformly for *all* $w \in \mathcal{W}$ under $P_{\boldsymbol{Z}}$. As a consequence, these bounds depend on structural properties of the hypothesis class $\mathcal{W}$ rather than on properties of the algorithm, and tend to be crude when applied to modern machine learning algorithms [6].

Prior Work: By generalizing a result obtained in [7] in the context of adaptive data analysis, Xu and Raginsky [8] obtained a bound on the average generalization error in terms of the mutual information $I(W; \boldsymbol{Z})$ between the output hypothesis W and the training data $\boldsymbol{Z}$. A drawback of the bound in [8] is that it is vacuous whenever the joint distribution $P_{W\boldsymbol{Z}}$ is not absolutely continuous with respect to $P_W P_{\boldsymbol{Z}}$, the product of the marginal distributions of W and $\boldsymbol{Z}$. This occurs, for example, when W is given by a deterministic function of $\boldsymbol{Z}$, and W and $\boldsymbol{Z}$ are separately continuous random variables. In [9], Bu *et al.* rectified this by obtaining a tighter bound in terms of the individual-sample mutual information $I(W; Z_i)$, which can be bounded even when $I(W; \boldsymbol{Z}) = \infty$. In [10], Asadi *et al.* combined the mutual information bound with the chaining technique [11], which exploits structure in the hypothesis class to tighten bounds. In some cases, this is shown to give stronger bounds than either the mutual information bound or the chaining bound individually.

To be evaluated, all of the aforementioned bounds require knowledge of the marginal distribution P_W, which depends

This work was partly supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Parts of the material of this paper were presented at the International Symposium on Information Theory (ISIT), June 2020, Los Angeles, CA [1]

F. Hellström and G. Durisi are with the Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden, (e-mail: {frehells,durisi}@chalmers.se).

on the data distribution P_Z. In practice, this data distribution is unknown, making the marginal P_W intractable. In light of this, Achille and Soatto [12] provided an upper bound on the mutual information between the training data and the output hypothesis in terms of the relative entropy between $P_{W|\boldsymbol{Z}}$ and a fixed, auxiliary distribution on the hypothesis space $\mathcal{W}$, and showed that this results in a computable upper bound on the average generalization error. Similarly, Negrea *et al.* [13] provided generalization bounds in terms of an auxiliary, possibly data-dependent distribution on $\mathcal{W}$. This weakens the bound, but makes it computable. Their use of the expected square root of the relative entropy $D(P_{W|\boldsymbol{Z}} \,||\, P_W)$, which they call *disintegrated* mutual information, in place of the mutual information leads to further improvements on the basic bound.

Recent studies, starting with the work of Steinke and Zakynthinou [14], have considered a setting with more structure, where it is assumed that a set $\widetilde{\boldsymbol{Z}}$ consisting of $2n$ independent and identically distributed (i.i.d.) training samples from P_Z is available, and that $\boldsymbol{Z}$ is formed by selecting n entries of $\widetilde{\boldsymbol{Z}}$ at random. We will refer to this setup as the *random-subset setting*, and call the setting without this additional structure the *standard setting*. In the random-subset setting, the average generalization error can be bounded by a quantity that depends on the conditional mutual information between the output hypothesis and the random variable that determines the selected training data $\boldsymbol{Z}$, given $\widetilde{\boldsymbol{Z}}$ [14, Thm. 5.1]. One advantage of this bound over the standard mutual information bound is that the conditional mutual information is always bounded. This broadens the applicability of the bound and results in tighter estimates. Also, as discussed in [14, Sec. 4], the conditional mutual information that appears in the bound has strong connections to classical generalization measures, such as VC dimension, compressibility, and stability. In [15], Haghifam *et al.* provided an individual-sample strengthening of this result, as well as improvements through their use of disintegration. In all of these derivations, the loss function is required to be bounded, which is a stronger requirement than what is needed in the standard setting.

All of the information-theoretic results discussed so far pertain to bounds on the average generalization error. In [16, App. A.3], Bassily *et al.* provided a PAC-Bayesian bound in terms of mutual information. This bound is essentially a data-independent relaxation of a well-known data-dependent bound from the PAC-Bayesian literature [17]. The dependence of the original data-dependent PAC-Bayesian bound on the confidence parameter δ is of order $\log(1/\delta)$, making it a high-probability bound. However, its mutual information relaxation in [16] has a less benign $1/\delta$-dependence. PAC-Bayesian techniques have recently found some success in producing non-vacuous generalization bounds for (randomized) deep neural networks. In [18], Dziugaite and Roy optimized a PAC-Bayesian bound to get non-vacuous generalization estimates for a simple neural network setup. These estimates were recently further improved in [19]. In [20], Zhou *et al.* derived a bound for compressed networks, i.e., small neural networks that are formed by pruning larger ones, and illustrated numerically that the bound is non-trivial for realistic settings. An extensive survey of the vast PAC-Bayesian literature, which is beyond the scope of this paper, can be found in, e.g., [3].

Finally, we survey the single-draw bounds that are relevant for our discussion. In addition to the aforementioned average and PAC-Bayesian bounds, both Xu and Raginsky [8, Thm. 3] and Bassily *et al.* [16] also provided single-draw generalization bounds in terms of mutual information. For both of them, the dependence on δ is of order $1/\delta$. In [21], Esposito *et al.* provided bounds in terms of a whole host of information-theoretic quantities, such as the Rényi divergence, the α-mutual information, and the maximal leakage. An interesting aspect of their α-mutual information bound is that, unlike the mutual information bounds in [8, Thm. 3] and [16], it is a high-probability bound. However, this bound does not imply a stronger mutual information bound. Indeed, if one lets $\alpha \to 1$, for which the α-mutual information reduces to the ordinary mutual information, the bound becomes vacuous. Bounds on the average generalization error are also provided [21, Sec. III.D], but these are generally weaker than the mutual information bounds in [8]. In the same vein, Dwork *et al.* derived single-draw generalization bounds in terms of other algorithmic stability measures, such as differential privacy [22] and (approximate) max-information [23]. These bounds are of the high-probability variety, but are typically weaker than the aforementioned maximal leakage bound [21, Sec. V]. All of the single-draw bounds mentioned here are data-independent.

Contributions: In this paper, we derive bounds of all three flavors—average, PAC-Bayesian, and single-draw—for both the standard setting and the random-subset setting. In the standard setting, we use the sub-Gaussianity of the loss function, together with a change of measure argument, to obtain an exponential inequality in terms of the *information density* between the hypothesis W and the training data $\boldsymbol{Z}$. This exponential inequality provides a framework that can be used not only to derive novel bounds, but also to recover several known results, which were originally derived using a host of different tools. In this sense, it provides a unifying approach for deriving information-theoretic generalization bounds. Through simple manipulations of the exponential inequality, we recover the average generalization bound in [8, Thm. 1] and the data-dependent PAC-Bayesian bound in [17, Prop. 3]. We also derive a novel data-dependent single-draw bound. Moreover, by further relaxing the PAC-Bayesian bound and the single-draw bound, we obtain two novel data-independent bounds that are explicit in the tth moments of the relative entropy $D(P_{W|\boldsymbol{Z}} \,||\, P_W)$ and of the information density, respectively. The dependence of these bounds on the confidence parameter δ is of order $1/\delta^t$. This is more favorable than that of similar bounds reported in [8] and [16], which have a dependence of order $1/\delta$. The moment bounds that we obtain illustrate that tighter estimates of the generalization error are available with higher confidence if the higher moments of the information measures that the bounds depend on are sufficiently small. Through a more refined analysis, we also obtain a high-probability data-independent single-draw bound in terms of maximal leakage. This result coincides with [21, Cor. 10], up to a logarithmic term. Finally, by using a different approach that relies on tools from binary hypothesis testing, we obtain a data-independent single-draw bound in terms of the tail of the information density. Similarly to the moment bounds, this bound illustrates that the faster the decay of the tail of the information density random variable, the more benign the dependence of the bound on δ.

Moving to the random-subset setting, we establish an exponential inequality, similar to that for the standard setting, in terms of the *conditional information density* between the hypothesis and a random variable that selects the data to be used for training, given all data samples. This exponential inequality is derived under the more stringent assumption of a bounded loss function. Then, we use this inequality to reobtain the average generalization bound in [14, Cor. 5.2], and to derive novel PAC-Bayesian and single-draw bounds, both of data-dependent and of data-independent flavor. Similarly to the standard setting, we also obtain a bound that is explicit in the tail of the conditional information density by using tools from binary hypothesis testing. Finally, inspired by [21], we derive a parametric inequality that can be used to obtain data-independent single-draw bounds. Using this inequality, we extend the results in [21] for bounded loss functions to the random-subset setting, and obtain bounds in terms of the conditional versions of the α-mutual information, the Rényi divergence, and the maximal leakage. Under some conditions, the conditional maximal leakage bound turns out to be stronger than its maximal leakage counterpart.

II. Preliminaries

In this section, we introduce some notation, define relevant information-theoretic quantities, and present some general results that will be used repeatedly in the remainder of this paper.

Standard and Random-Subset Settings: Let $\mathcal{Z}$ be the instance space, $\mathcal{W}$ be the hypothesis space, and $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ be the loss function. In the standard setting, n training samples $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ are available. These n samples constitute the training data. We assume that all entries of $\boldsymbol{Z}$ are drawn independently from some unknown distribution P_Z on $\mathcal{Z}$. In the random-subset setting, $2n$ training samples $\widetilde{\boldsymbol{Z}} = (\tilde{Z}_1, \ldots, \tilde{Z}_{2n})$ are available, with all entries of $\widetilde{\boldsymbol{Z}}$ being drawn independently from P_Z. However, only a randomly selected subset of cardinality n is actually used as the training data. Following [14], we assume that the training data $\boldsymbol{Z}(\boldsymbol{S})$ is selected as follows. Let $\boldsymbol{S} = (S_1, \ldots, S_n)$ be an n-dimensional random vector, the elements of which are drawn independently from a $\mathrm{Bern}(1/2)$ distribution and are independent of $\widetilde{\boldsymbol{Z}}$. Then, for $i = 1, \ldots, n$, the ith training sample in $\boldsymbol{Z}(\boldsymbol{S})$ is $Z_i(S_i) = \tilde{Z}_{i+S_i n}$. A randomized learning algorithm is a conditional distribution $P_{W|\boldsymbol{Z}}$. We let $L_{\boldsymbol{Z}}(W) = \frac{1}{n}\sum_{i=1}^n \ell(W, Z_i)$ denote the empirical loss and $L_{P_Z}(W) = \mathbb{E}_{P_Z}[\ell(W, Z)]$ the population loss. The generalization error is defined as $\mathrm{gen}(W, \boldsymbol{Z}) = L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)$.

Information Measures: A quantity that will appear in many of our bounds is the information density, defined as

$$\imath(W, \boldsymbol{Z}) = \log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}} \tag{1}$$

where $\mathrm{d}P_{W\boldsymbol{Z}}/\,\mathrm{d}P_W P_{\boldsymbol{Z}}$ is the Radon-Nikodym derivative of $P_{W\boldsymbol{Z}}$ with respect to $P_W P_{\boldsymbol{Z}}$. Here, P_W is the distribution induced on the hypothesis space $\mathcal{W}$ by P_Z through $P_{W|\boldsymbol{Z}}$. The information density is well defined whenever $P_{W\boldsymbol{Z}}$ is absolutely continuous with respect to $P_W P_{\boldsymbol{Z}}$, which we denote by $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$. The name information density is motivated by the fact that its expectation under $P_{W\boldsymbol{Z}}$ is the mutual information $I(W; \boldsymbol{Z})$. In the random-subset setting, several of our bounds will be in terms of the conditional information density

$$\imath(W, \boldsymbol{S}|\widetilde{\boldsymbol{Z}}) = \log \frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}} P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}} \tag{2}$$

where $P_{W|\widetilde{\boldsymbol{Z}}}$ is a conditional distribution on $\mathcal{W}$ given $\widetilde{\boldsymbol{Z}}$, obtained by marginalizing out $\boldsymbol{S}$. Here, the absolute continuity requirement is that $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}} \ll P_{W|\widetilde{\boldsymbol{Z}}} P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$. In the random-subset setting, this is satisfied since $P_{W|\widetilde{\boldsymbol{Z}}}$ is obtained by marginalizing out the discrete random variable $\boldsymbol{S}$ from $P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$. If we take the expectation of $\imath(W, \boldsymbol{S}|\widetilde{\boldsymbol{Z}})$ under the joint distribution $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$, we obtain the conditional mutual information $I(W; \boldsymbol{S}|\widetilde{\boldsymbol{Z}})$, a key quantity in the bounds developed in [14].

Let $\alpha \in (0, 1) \cup (1, \infty)$. The Rényi divergence of order α is defined as [24]

$$(\alpha - 1) D_\alpha(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) = \log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[e^{\alpha \imath(W, \boldsymbol{Z})}\right]. \tag{3}$$

In the limit $\alpha \to 1$, it reduces to the relative entropy $D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}})$. The conditional Rényi divergence of order α is given by [25]

$$\begin{aligned}(\alpha - 1) D_\alpha(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}} P_{\boldsymbol{S}|\widetilde{\boldsymbol{Z}}} \,||\, P_{W|\widetilde{\boldsymbol{Z}}} P_{\boldsymbol{S}|\widetilde{\boldsymbol{Z}}} \,|\, P_{\widetilde{\boldsymbol{Z}}}) \\ = \log \mathbb{E}_{P_{\widetilde{\boldsymbol{Z}}} P_{W|\widetilde{\boldsymbol{Z}}} P_{\boldsymbol{S}|\widetilde{\boldsymbol{Z}}}}\left[\exp\Big(\alpha \imath(W, \boldsymbol{S}|\widetilde{\boldsymbol{Z}})\Big)\right].\end{aligned} \tag{4}$$

The α-mutual information, which is studied in depth in [25], is defined as

$$(\alpha - 1) I_\alpha(\boldsymbol{Z}; W) = \log \mathbb{E}^\alpha_{P_W}\left[\mathbb{E}^{1/\alpha}_{P_{\boldsymbol{Z}}}[\exp(\alpha \imath(W, \boldsymbol{Z}))]\right]. \tag{5}$$

In the limit $\alpha \to 1$, it reduces to the mutual information $I(W; \boldsymbol{Z})$, whereas for $\alpha \to \infty$, it becomes the maximal leakage [26]:

$$\mathcal{L}(\boldsymbol{Z} \to W) = \log \mathbb{E}_{P_W}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{Z}}} \exp(\imath(W, \boldsymbol{Z}))\right]. \tag{6}$$

Here, the essential supremum of a measurable function $f(\cdot)$ of a random variable $\boldsymbol{Z}$ distributed as $P_{\boldsymbol{Z}}$ is defined as

$$\operatorname*{ess\,sup}_{P_{\boldsymbol{Z}}} f(\boldsymbol{Z}) = \inf_{a \in \mathbb{R}}\Big[P_{\boldsymbol{Z}}\big(\{\boldsymbol{Z} : f(\boldsymbol{Z}) > a\}\big) = 0\Big]. \tag{7}$$

The conditional α-mutual information does not have a commonly accepted definition. In [27], three definitions are provided and given operational interpretations, two of which have known closed-form expressions. The first coincides with the conditional Rényi divergence, while the second, which we will term $I_\alpha(W; \boldsymbol{S}|\widetilde{\boldsymbol{Z}})$, is defined as

$$\begin{aligned}(\alpha - 1) I_\alpha(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \\ = \log \mathbb{E}_{P_{\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^\alpha_{P_{W|\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\alpha}_{P_{\boldsymbol{S}|\tilde{\boldsymbol{Z}}}}\left[\exp\Big(\alpha \imath(W, \boldsymbol{S}|\widetilde{\boldsymbol{Z}})\Big)\right]\right]\right].\end{aligned} \tag{8}$$

In the limit $\alpha \to \infty$, this reduces to the conditional maximal leakage [26, Thm. 6]

$$\mathcal{L}(\boldsymbol{S} \to W|\widetilde{\boldsymbol{Z}}) = \log \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}} \mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}}}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}|\tilde{\boldsymbol{Z}}}} e^{\imath(W, \boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]. \tag{9}$$

Note that $\boldsymbol{S}$ and $\widetilde{\boldsymbol{Z}}$ are independent in the random-subset setting. Hence, $P_{\boldsymbol{S}|\widetilde{\boldsymbol{Z}}}$ can be replaced by $P_{\boldsymbol{S}}$ in (4), (8), and (9).

Useful Results: Many previous studies have used the data-processing inequality as a tool for deriving generalization bounds [16], [21]. In binary hypothesis testing, it is known that the data-processing inequality only provides weak converse bounds on the region of achievable error rates. To get strong converse bounds, one relies on the following lemma instead [28, Lem. 12.2].

Lemma 1 (Strong Converse Lemma): Let P and Q be probability distributions on some common space $\mathcal{X}$ such that P is absolutely continuous with respect to Q, and let $\mathcal{E} \in \mathcal{X}$ be a measurable set. Then, for all $\gamma \in \mathbb{R}$,

$$P[\mathcal{E}] \leq P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma\right] + e^{\gamma} Q[\mathcal{E}]. \tag{10}$$

In Section III-C2 and Section IV-C2, we will show how to use this result to derive generalization bounds.

We will also make repeated use of the following result, due to Hoeffding [29, Prop. 2.5].

Lemma 2 (Hoeffding's Inequality): Let $X \sim P_X$ be a σ-sub-Gaussian random variable, i.e., a random variable satisfying the following inequality for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right). \tag{11}$$

Then, for all $\epsilon > 0$,

$$P_X(|X - \mathbb{E}[X]| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{12}$$

Note that a random variable bounded on $[a, b]$ is σ-sub-Gaussian with $\sigma = (b-a)/2$. Also, if X_i, for $i = 1, \ldots, n$, are independent σ-sub-Gaussian random variables, the average $(1/n)\sum_{i=1}^{n} X_i$ is $\sigma/\sqrt{n}$-sub-Gaussian.

III. Generalization Bounds for the Standard Setting

In this section, we study the standard setting described in Section II. We will assume that the loss function $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$. This means that, for all $\lambda \in \mathbb{R}$ and for all $w \in \mathcal{W}$,

$$\mathbb{E}_{P_Z}[\exp(\lambda(\mathbb{E}_{P_Z}[\ell(w, Z)] - \ell(w, Z))] \leq e^{\lambda^2 \sigma^2 / 2}. \tag{13}$$

We will derive bounds on the generalization error of a probabilistic learning algorithm $P_{W|\boldsymbol{Z}}$ in terms of some function of the information density (1). As previously mentioned, several different notions of generalization error bounds have been investigated in the literature. One such notion is that of average generalization bounds, where we want to find an ϵ such that

$$|\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})]| \leq \epsilon. \tag{14}$$

This ϵ will in general depend on the joint distribution $P_{W\boldsymbol{Z}}$, on properties of the loss function, and on the cardinality n of the training data. We will study this type of bounds in Section III-A.

Another approach, typically studied in the PAC-Bayesian literature, is to find probabilistic bounds of the following form: with probability at least $1 - \delta$ under $P_{\boldsymbol{Z}}$,

$$\left|\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})]\right| \leq \epsilon. \tag{15}$$

This bound is interesting when we have a fixed data set $\boldsymbol{Z}$, but draw a new hypothesis according to $P_{W|\boldsymbol{Z}}$ each time we want to use our algorithm. The main advantage of this PAC-Bayesian approach is that, by considering a distribution over the hypothesis class rather than just a single hypothesis, one can capture uncertainty about the hypotheses and exploit possible correlations between them [17]. We derive bounds of this type in Section III-B.

Finally, we also consider the single-draw scenario. In this setting, we are interested in bounds of the following flavor: with probability at least $1 - \delta$ under $P_{W\boldsymbol{Z}}$,

$$|\mathrm{gen}(W, \boldsymbol{Z})| \leq \epsilon. \tag{16}$$

This type of result is relevant when we draw a single hypothesis W based on our training data, and want to bound the generalization error of this particular W with high probability. The probabilistic bounds in (15) and (16) are said to be high-probability bounds if the dependence of ϵ on the confidence parameter δ is at most of order $\log(1/\delta)$.

In Theorem 1 below, we present an exponential inequality that will be used in Section III-A, Section III-B, and Section III-C to derive generalization bounds of all three flavors. The derivation of this exponential inequality and its use to obtain generalization bounds draw inspiration from [4, Sec. 1.2.4], where a similar approach is used to obtain PAC-Bayesian and single-draw bounds for the special case in which the loss function has range restricted to $\{0, 1\}$.

Theorem 1: Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ consist of n i.i.d. training samples generated from P_Z, and let $P_{W|\boldsymbol{Z}}$ be a probabilistic learning algorithm. Assume that $\ell(w, Z) : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$. Also, assume that $P_{W\boldsymbol{Z}}$ is absolutely continuous with respect to $P_W P_{\boldsymbol{Z}}$. Then, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\exp\left(\lambda \mathrm{gen}(W, \boldsymbol{Z}) - \frac{\lambda^2 \sigma^2}{2n} - \imath(W, \boldsymbol{Z})\right)\right] \leq 1. \tag{17}$$

Proof: Since $\ell(w, Z)$ is σ-sub-Gaussian for all $w \in \mathcal{W}$ and the Z_i are i.i.d., $(1/n)\sum_{i=1}^{n} \ell(w, Z_i)$ is $\sigma/\sqrt{n}$-sub-Gaussian for all $w \in \mathcal{W}$, as remarked after Lemma 2. Thus, for all $w \in \mathcal{W}$,

$$\mathbb{E}_{P_{\boldsymbol{Z}}}\left[\exp\left(\lambda\left(\mathbb{E}_{P_Z}[\ell(w, Z)] - \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)\right)\right)\right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2n}\right). \tag{18}$$

Reorganizing terms and taking the expectation with respect to P_W, we get

$$\mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[\exp\left(\lambda \mathrm{gen}(W, \boldsymbol{Z}) - \frac{\lambda^2 \sigma^2}{2n}\right)\right] \leq 1. \tag{19}$$

Now, let $\mathcal{E} = \mathrm{supp}(P_{W\boldsymbol{Z}})$ be the support of $P_{W\boldsymbol{Z}}$. From (19), it follows that

$$\mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[1_{\mathcal{E}} \cdot \exp\left(\lambda \mathrm{gen}(W, \boldsymbol{Z}) - \frac{\lambda^2 \sigma^2}{2n}\right)\right] \leq 1 \tag{20}$$

where $1_{\mathcal{E}}$ is the indicator function of the set $\mathcal{E}$. To obtain (17), we perform a change of measure from $P_W P_{\boldsymbol{Z}}$ to $P_{W\boldsymbol{Z}}$ [28, Prop. 17.1]. ∎

Note that Theorem 1 holds *verbatim* if P_W is replaced with an auxiliary distribution Q_W, under a suitable absolute continuity assumption. This is detailed in the next remark.

Remark 1: Consider the setting of Theorem 1, but with the altered absolute continuity assumption that $P_{W\boldsymbol{Z}} \ll Q_W P_{\boldsymbol{Z}}$ for some distribution Q_W on $\mathcal{W}$. Then,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\exp\left(\lambda \mathrm{gen}(W,\boldsymbol{Z}) - \frac{\lambda^2\sigma^2}{2n} - \log\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}Q_W P_{\boldsymbol{Z}}}\right)\right] \leq 1. \quad (21)$$

For the bounds that we will later derive, the choice $Q_W = P_W$ is optimal. Unfortunately, since the data distribution P_Z is considered to be unknown in the statistical learning framework, the marginal distribution P_W is also unavailable. Hence, P_W needs to be replaced by some suitably chosen auxiliary distribution Q_W whenever one wants to numerically evaluate the generalization bounds that we derive later in this section.[1] In the remainder of this paper, all bounds will be given in terms of P_W. Thanks to this choice, many of the terms that appear in our results will be expressible in terms of familiar information-theoretic quantities. However, through repeated references to Remark 1, we will emphasize that the bounds can easily be generalized to the case in which P_W is replaced by an auxiliary distribution Q_W.

A. Average Generalization Error Bounds

We now use Theorem 1 to obtain an average generalization error bound of the form given in (14).

Corollary 1: Under the setting of Theorem 1,

$$|\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})]| \leq \sqrt{\frac{2\sigma^2}{n} I(W;\boldsymbol{Z})}. \quad (22)$$

Proof: We apply Jensen's inequality to (17), resulting in

$$\exp\Bigg(\lambda\,\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})] - \frac{\lambda^2\sigma^2}{2n} - \mathbb{E}_{P_{W\boldsymbol{Z}}}[\imath(W,\boldsymbol{Z})]\Bigg) \leq 1. \quad (23)$$

Note that $\mathbb{E}_{P_{W\boldsymbol{Z}}}[\imath(W,\boldsymbol{Z})] = D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) = I(W;\boldsymbol{Z})$. By reorganizing terms, we get

$$\frac{\lambda^2\sigma^2}{2n} - \lambda\,\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})] + D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) \geq 0. \quad (24)$$

Since (24) gives a nonnegative parabola in λ, its discriminant must be nonpositive. We thereby obtain

$$\mathbb{E}^2_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})] - \frac{2\sigma^2}{n} D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) \leq 0 \quad (25)$$

from which (22) directly follows. ∎

The bound in (22) coincides with the result reported in [8, Thm. 1]. As noted in Remark 1, we can substitute an arbitrary Q_W for P_W in (25), provided that the necessary absolute continuity criterion is fulfilled. This leads to a more general bound involving the relative entropy $D(P_{W\boldsymbol{Z}} \,||\, Q_W P_{\boldsymbol{Z}})$.

[1]This issue is well understood in the PAC-Bayesian literature, where the available bounds are given in terms of an auxiliary (*prior*) distribution Q_W that does not depend on the unknown data distribution P_Z.

B. PAC-Bayesian Generalization Error Bounds

We now turn to PAC-Bayesian bounds of the form given in (15). In the following corollary, we reobtain a known data-dependent bound [17, Prop. 3] and present a novel data-independent relaxation.

Corollary 2: Under the setting of Theorem 1, the following holds with probability at least $1-\delta$ under $P_{\boldsymbol{Z}}$ for all $t > 0$:

$$\big|\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})]\big| \leq \sqrt{\frac{2\sigma^2}{n}\left(D(P_{W|\boldsymbol{Z}} \,||\, P_W) + \log\frac{1}{\delta}\right)} \quad (26)$$

$$\leq \sqrt{\frac{2\sigma^2}{n}\left(\frac{\mathbb{E}^{1/t}_{P_{\boldsymbol{Z}}}\big[D(P_{W|\boldsymbol{Z}} \,||\, P_W)^t\big]}{(\delta/2)^{1/t}} + \log\frac{2}{\delta}\right)}. \quad (27)$$

Here, the first inequality yields a data-dependent bound, while the second inequality provides a data-independent relaxation.

Proof: As in Corollary 1, we start from (17) and use Jensen's inequality, but now only with respect to $P_{W|\boldsymbol{Z}}$. This leads to

$$\mathbb{E}_{P_{\boldsymbol{Z}}}\Bigg[\exp\Bigg(\lambda\,\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})] - \frac{\lambda^2\sigma^2}{2n} - D(P_{W|\boldsymbol{Z}} \,||\, P_W)\Bigg)\Bigg] \leq 1 \quad (28)$$

where we used that, for a fixed $\boldsymbol{Z}$,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\imath(W,\boldsymbol{Z})] = D(P_{W|\boldsymbol{Z}} \,||\, P_W). \quad (29)$$

Next, we use the following result. Let $U \sim P_U$ be a nonnegative random variable satisfying $\mathbb{E}[U] \leq 1$. Then, Markov's inequality implies that

$$P_U[U \leq 1/\delta] \geq 1 - \mathbb{E}[U]\,\delta \geq 1 - \delta. \quad (30)$$

By applying (30) to the random variable in (28), we obtain

$$P_{\boldsymbol{Z}}\Bigg[\exp\Bigg(\lambda\,\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})] - \frac{\lambda^2\sigma^2}{2n} - D(P_{W|\boldsymbol{Z}} \,||\, P_W)\Bigg) \leq \frac{1}{\delta}\Bigg] \geq 1 - \delta. \quad (31)$$

Reorganizing terms, we conclude that

$$P_{\boldsymbol{Z}}\Bigg[\frac{\lambda^2\sigma^2}{2n} - \lambda\,\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})] + D(P_{W|\boldsymbol{Z}} \,||\, P_W) + \log\frac{1}{\delta} \geq 0\Bigg] \geq 1 - \delta. \quad (32)$$

By rearranging the inequality inside the probability in (32), we obtain a nonnegative parabola in λ. Using the nonpositivity of its discriminant, as in (25), we arrive at (26). To prove (27), we apply Markov's inequality to the random variable $D(P_{W|\boldsymbol{Z}} \,||\, P_W)^t$, which after some manipulation yields

$$P_{\boldsymbol{Z}}\left[D(P_{W|\boldsymbol{Z}} \,||\, P_W) \leq \frac{\mathbb{E}^{1/t}_{P_{\boldsymbol{Z}}}\big[D(P_{W|\boldsymbol{Z}} \,||\, P_W)^t\big]}{\delta^{1/t}}\right] \geq 1 - \delta. \quad (33)$$

We now use the union bound to combine (26) with (33) and perform the substitution $\delta \to \delta/2$, after which (27) follows. ∎

Note that by setting the parameter $t = 1$ in (27), we get $\mathbb{E}_{P_{\boldsymbol{Z}}}\big[D(P_{W|\boldsymbol{Z}} \,||\, P_W)\big] = I(W;\boldsymbol{Z})$. This choice of t recovers the result reported in [16, App. 3]. Instead, if we let $t \to \infty$, the polynomial δ-dependence in (27) disappears and the bound becomes a high-probability bound. This illustrates that one can get progressively better dependence on δ by letting the bound depend on higher moments of $D(P_{W|\boldsymbol{Z}} \,||\, P_W)$. The tightness of the resulting bound depends on how well one can control these higher moments. Finally, as per Remark 1, we can obtain more general bounds by replacing P_W in (26) and (27) with an arbitrary Q_W that satisfies a suitable absolute continuity property.

C. Single-Draw Generalization Error Bounds

We now turn our attention to single-draw bounds of the form given in (16). We will derive generalization bounds by using two different approaches. Our first approach relies on the exponential inequality from Theorem 1, which we use to get a data-dependent bound in terms of the information density $\imath(W,\boldsymbol{Z})$. We then relax this result in different ways to obtain several data-independent bounds. Our second approach, which yields a generalization bound that is explicit in the tail of the information density, relies on the change of measure result stated in Lemma 1. This bound can be relaxed to obtain essentially the same data-independent bounds obtained using the first approach.

1) Generalization Bounds from the Exponential Inequality: We begin by using Theorem 1 to derive a data-dependent single-draw generalization bound and a data-independent relaxation, similar to the PAC-Bayesian results in Corollary 2. Both of these bounds are novel.

Corollary 3: Under the setting of Theorem 1, with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$, the following inequalities hold for all $t > 0$:[2]

$$|\mathrm{gen}(W,\boldsymbol{Z})| \le \sqrt{\frac{2\sigma^2}{n}\left(\imath(W,\boldsymbol{Z}) + \log\frac{1}{\delta}\right)} \tag{34}$$

$$\le \sqrt{\frac{2\sigma^2}{n}\left(I(W;\boldsymbol{Z}) + \frac{M_t(W;\boldsymbol{Z})}{(\delta/2)^{1/t}} + \log\frac{2}{\delta}\right)}. \tag{35}$$

Here, the first inequality provides a data-dependent bound and the second inequality is a data-independent relaxation. In (35), $M_t(W;\boldsymbol{Z})$ is the tth root of the tth central moment of $\imath(W,\boldsymbol{Z})$:

$$M_t(W;\boldsymbol{Z}) = \mathbb{E}^{1/t}_{P_{W\boldsymbol{Z}}}\Big[\big|\imath(W,\boldsymbol{Z}) - D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}})\big|^t\Big]. \tag{36}$$

Proof: By directly applying Markov's inequality (30) to (17), we conclude that

$$P_{W\boldsymbol{Z}}\left[\exp\left(\lambda\,\mathrm{gen}(W,\boldsymbol{Z}) - \frac{\lambda^2\sigma^2}{2n} - \imath(W,\boldsymbol{Z})\right) \le \frac{1}{\delta}\right] \ge 1-\delta. \tag{37}$$

Rearranging the inequality inside the probability in (37) yields a nonnegative parabola in λ. By using the nonpositivity of its discriminant, as was done after (32), we arrive at (34). To prove (35), we use Markov's inequality in the following form: for a random variable $U \sim P_U$, the following holds for all $t > 0$:

$$P_U\left[U \le \mathbb{E}[U] + \frac{\mathbb{E}^{1/t}[|U - \mathbb{E}[U]|^t]}{\delta^{1/t}}\right] \ge 1-\delta. \tag{38}$$

Applying (38) with $U = \imath(W,\boldsymbol{Z})$ and using the union bound to combine the resulting inequality with (34), we obtain (35) after performing the substitution $\delta \to \delta/2$. ∎

As usual, we can obtain more general bounds by substituting Q_W for P_W in Corollary 3, provided that the necessary absolute continuity assumption is satisfied.

Similarly to what we noted for the PAC-Bayesian bound (27), the δ-dependence in (35) can be made more benign by letting the bound depend on higher central moments of $\imath(W,\boldsymbol{Z})$, but the tightness of the resulting bound hinges on how well one can control these higher moments. In particular, if we let $t \to \infty$ in (35), we obtain the following high-probability bound:

$$|\mathrm{gen}(W,\boldsymbol{Z})| \le \sqrt{\frac{2\sigma^2}{n}\left(I(W;\boldsymbol{Z}) + M_\infty(W;\boldsymbol{Z}) + \log\frac{2}{\delta}\right)}. \tag{39}$$

Here, $M_\infty(W;\boldsymbol{Z})$ is given by

$$M_\infty(W;\boldsymbol{Z}) = \operatorname*{ess\,sup}_{P_{W\boldsymbol{Z}}} \big|\imath(W,\boldsymbol{Z}) - I(W;\boldsymbol{Z})\big|. \tag{40}$$

Note that the supremization in (40) is over the argument of $\imath(W,\boldsymbol{Z})$, whereas $I(W;\boldsymbol{Z})$ is a constant.

The data-independent relaxation in Corollary 3 is not as tight as the one obtained in Corollary 2. Indeed, since $\imath(W,\boldsymbol{Z})$ can be negative, we had to use a weaker version of Markov's inequality (compare (38) with (30)). In the following corollary, we provide two alternative data-independent bounds. The first bound depends on the maximal leakage $\mathcal{L}(\boldsymbol{Z} \to W)$ defined in (6), and recovers [21, Cor. 10] up to a logarithmic term. The second bound, which is novel, is in terms of the Rényi divergence (3).

Corollary 4: Under the setting of Theorem 1, the following inequalities hold with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$:

$$|\mathrm{gen}(W,\boldsymbol{Z})| \le \sqrt{\frac{2\sigma^2}{n}\left(\mathcal{L}(\boldsymbol{Z} \to W) + 2\log\frac{2}{\delta}\right)} \tag{41}$$

and, for all $\alpha, \gamma > 1$ such that $1/\alpha + 1/\gamma = 1$,

$$|\mathrm{gen}(W,\boldsymbol{Z})| \le \left[\frac{2\sigma^2}{n}\left(\frac{\alpha-1}{\alpha}D_\alpha(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) + \frac{\gamma-1}{\gamma}D_\gamma(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) + 2\log\frac{2}{\delta}\right)\right]^{1/2}. \tag{42}$$

Proof: By applying Markov's inequality, we conclude that with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$,

$$\imath(W,\boldsymbol{Z}) \le \log \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right] + \log\left(\frac{1}{\delta}\right). \tag{43}$$

Next, the expectation over $P_{\boldsymbol{Z}|W}$ can be replaced by an essential supremum to obtain the inequality

$$\mathbb{E}_{P_W P_{\boldsymbol{Z}|W}}\left[\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right] \le \mathbb{E}_{P_W}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{Z}|W}} \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right]. \tag{44}$$

[2]Note that the argument of the square root in (34) can be negative, but that this happens with probability at most δ. Therefore, the right-hand side of (34) is well-defined with probability at least $1-\delta$.

The assumption that $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$ means that any set in the support of $P_{W\boldsymbol{Z}}$ is also in the support of $P_W P_{\boldsymbol{Z}}$. We can therefore upper-bound the $\operatorname{ess\,sup}$ as

$$\operatorname*{ess\,sup}_{P_{\boldsymbol{Z}|W}} \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}} \leq \operatorname*{ess\,sup}_{P_{\boldsymbol{Z}}} \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}. \tag{45}$$

By using the union bound to combine (43)-(45) with (34) and performing the substitution $\delta \to \delta/2$, we obtain (41).

To prove (42), we first apply Markov's inequality and then perform a change of measure to conclude that the following inequalities hold with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$:

$$\imath(W,\boldsymbol{Z}) \leq \log \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right] + \log\frac{1}{\delta} \tag{46}$$

$$\leq \log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[\left(\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right)^2\right] + \log\frac{1}{\delta}. \tag{47}$$

Next, we apply Hölder's inequality twice as follows. Let $\alpha, \gamma, \alpha', \gamma' > 1$ be constants chosen so as to satisfy $1/\alpha + 1/\gamma = 1/\alpha' + 1/\gamma' = 1$. Then,

$$\begin{aligned}
&\mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[\left(\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right)^2\right] \\
&\leq \mathbb{E}_{P_W}\left[\mathbb{E}^{1/\alpha}_{P_{\boldsymbol{Z}}}\left[e^{\alpha\imath(W,\boldsymbol{Z})}\right]\cdot \mathbb{E}^{1/\gamma}_{P_{\boldsymbol{Z}}}\left[e^{\gamma\imath(W,\boldsymbol{Z})}\right]\right] &(48)\\
&\leq \mathbb{E}^{1/\alpha'}_{P_W}\left[\mathbb{E}^{\alpha'/\alpha}_{P_{\boldsymbol{Z}}}\left[e^{\alpha\imath(W,\boldsymbol{Z})}\right]\right]\cdot \mathbb{E}^{1/\gamma'}_{P_W}\left[\mathbb{E}^{\gamma'/\gamma}_{P_{\boldsymbol{Z}}}\left[e^{\gamma\imath(W,\boldsymbol{Z})}\right]\right]. &(49)
\end{aligned}$$

Setting $\alpha = \alpha'$, which implies $\gamma = \gamma'$, we conclude that

$$\begin{aligned}
&\log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[\left(\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right)^2\right] \\
&\leq \log \mathbb{E}^{1/\alpha}_{P_W P_{\boldsymbol{Z}}}\left[e^{\alpha\imath(W,\boldsymbol{Z})}\right] + \log \mathbb{E}^{1/\gamma}_{P_W P_{\boldsymbol{Z}}}\left[e^{\gamma\imath(W,\boldsymbol{Z})}\right] &(50)\\
&= \frac{\alpha-1}{\alpha} D_\alpha(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) + \frac{\gamma-1}{\gamma} D_\gamma(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}). &(51)
\end{aligned}$$

Substituting (51) into (47), and then combining (47) with (34) through the union bound, we establish (42) after the substitution $\delta \to \delta/2$. ∎

The bound in (41) coincides with the maximal leakage bound in [21, Cor. 10], up to a $(2\sigma^2/n)\log(2/\delta)$ term inside the square root. It is stronger than the max information bound in [23, Thm. 4], for the case in which the parameter β therein is set to 0, and also stronger than (39), up to the same logarithmic term. Indeed, let the max information be defined as

$$I_{\max}(W;\boldsymbol{Z}) = \operatorname*{ess\,sup}_{P_{W\boldsymbol{Z}}} \imath(W,\boldsymbol{Z}). \tag{52}$$

As shown in [21, Lem. 12], we have $\mathcal{L}(\boldsymbol{Z}\to W) \leq I_{\max}(W;\boldsymbol{Z})$. It is also readily verified that

$$I_{\max}(W;\boldsymbol{Z}) \leq I(W;\boldsymbol{Z}) + M_\infty(W;\boldsymbol{Z}). \tag{53}$$

We thus have the chain of inequalities

$$\mathcal{L}(\boldsymbol{Z}\to W) \leq I_{\max}(W;\boldsymbol{Z}) \leq I(W;\boldsymbol{Z}) + M_\infty(W;\boldsymbol{Z}). \tag{54}$$

In particular, provided that

$$\mathcal{L}(\boldsymbol{Z}\to W) \leq I_{\max}(W;\boldsymbol{Z}) + \log\frac{2}{\delta} \tag{55}$$

the bound in (41) is tighter than the max information bound in [23, Thm. 4] with $\beta = 0$, and also tighter than (39). Still, the bound in [21, Cor. 10] is stronger due to the aforementioned logarithmic term.

As usual, we can obtain more general bounds by replacing P_W with an arbitrary Q_W in Corollary 4, provided that $P_{W\boldsymbol{Z}} \ll Q_W P_{\boldsymbol{Z}}$. However, for the proof of (41), we still need the original absolute continuity assumption $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$ to guarantee that (45) holds. Note that a similar extension can easily be performed on [21, Thm. 1] and on the corollaries that are based on it, including [21, Cor. 10].

2) Generalization Bounds from the Strong Converse: Next, we use Lemma 1 to derive an additional data-independent single-draw generalization bound. This novel bound depends on the tail of the information density.

Theorem 2: Under the setting of Theorem 1, with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$, the following holds:

$$\begin{aligned}
&|\mathrm{gen}(W,\boldsymbol{Z})| \leq \\
&\sqrt{\frac{2\sigma^2}{n}\left(\gamma + \log\left(\frac{2}{\delta - P_{W\boldsymbol{Z}}[\imath(W,\boldsymbol{Z}) \geq \gamma]}\right)\right)}.
\end{aligned} \tag{56}$$

This is valid for all γ such that the right-hand side is defined and real.

Proof: The proof relies on Lemma 1. We set $P = P_{W\boldsymbol{Z}}$, $Q = P_W P_{\boldsymbol{Z}}$, and

$$\mathcal{E} = \{W,\boldsymbol{Z} : |\mathrm{gen}(W,\boldsymbol{Z})| > \epsilon\}. \tag{57}$$

Due to the σ-sub-Gaussianity of the loss function, Hoeffding's inequality (Lemma 2) implies that

$$\begin{aligned}
P_W P_{\boldsymbol{Z}}[\mathcal{E}] &= P_W P_{\boldsymbol{Z}}[|L_{\boldsymbol{Z}}(W) - \mathbb{E}_{P_{\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)]| > \epsilon] &(58)\\
&\leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right). &(59)
\end{aligned}$$

Substituting (58) into (10), we get

$$\begin{aligned}
&P_{W\boldsymbol{Z}}[|\mathrm{gen}(W,\boldsymbol{Z})| > \epsilon] \\
&\qquad \leq P_{W\boldsymbol{Z}}[\imath(W,\boldsymbol{Z}) \geq \gamma] + 2\exp\left(\gamma - n\frac{\epsilon^2}{2\sigma^2}\right).
\end{aligned} \tag{60}$$

We obtain the desired result by requiring the right-hand side of (60) to equal δ and solving for ϵ. ∎

As for the previous results, a more general bound can be obtained by setting $Q = Q_W P_{\boldsymbol{Z}}$, where Q_W is an arbitrary auxiliary distribution on $\mathcal{W}$, provided that a suitable absolute continuity criterion is fulfilled.

The result in Theorem 2 indicates a trade-off between the decay of the tail of the information density and the tightness of the generalization bound. Indeed, the parameter γ has to be chosen sufficiently large to make the argument of the logarithm positive. However, increasing γ too much may yield a loose bound because of the γ term that is added to the logarithm.

The bound in Theorem 2 can be relaxed to recover some of the data-independent bounds discussed earlier in this section, up to a $(2\sigma^2/n)\log 2$ penalty term inside the square root. In Remarks 2 and 3, we present these alternative derivations.

Remark 2 (Alternative derivation of the moment bound (35)*):* Using Markov's inequality, we conclude that

$$P_{W\boldsymbol{Z}}[\imath(W,\boldsymbol{Z}) \geq \gamma] \leq P_{W\boldsymbol{Z}}\Big[|\imath(W,\boldsymbol{Z}) - D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}})| \geq \gamma - D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}})\Big] \tag{61}$$

$$\leq \frac{(M_t(W;\boldsymbol{Z}))^t}{(\gamma - D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}))^t} \tag{62}$$

where $M_t(W;\boldsymbol{Z})$ is defined in (36). Next, we set

$$\gamma = D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) + \frac{M_t(W;\boldsymbol{Z})}{(\delta/2)^{1/t}} \tag{63}$$

which, once it is substituted into (61), implies the inequality $P_{W\boldsymbol{Z}}[\imath(W,\boldsymbol{Z}) \geq \gamma] \leq \delta/2$. Using this in (56), we obtain

$$|\mathrm{gen}(W,\boldsymbol{Z})| \leq \left[\frac{2\sigma^2}{n}\left(D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) + \frac{M_t(W;\boldsymbol{Z})}{(\delta/2)^{1/t}} + \log\frac{4}{\delta}\right)\right]^{1/2}. \tag{64}$$

This coincides with the bound in (35), up to a $(2\sigma^2/n)\log 2$ term inside the square root.

Remark 3 (Alternative derivation of the maximal leakage bound (41)*):* Note that

$$P_{W\boldsymbol{Z}}[\imath(W,\boldsymbol{Z}) \geq \gamma] \leq P_W\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{Z}|W}} e^{\imath(W,\boldsymbol{Z})} \geq e^{\gamma}\right]. \tag{65}$$

Since $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$, the $\operatorname{ess\,sup}$ can be upper-bounded as in (45). Hence,

$$P_{W\boldsymbol{Z}}[\imath(W,\boldsymbol{Z}) \geq \gamma] \leq P_W\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{Z}}} e^{\imath(W,\boldsymbol{Z})} \geq e^{\gamma}\right]. \tag{66}$$

By applying Markov's inequality to the right-hand side of (66), we find that

$$P_{W\boldsymbol{Z}}[\imath(W,\boldsymbol{Z}) \geq \gamma] \leq e^{-\gamma}\,\mathbb{E}_{P_W}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{Z}}} e^{\imath(W,\boldsymbol{Z})}\right] = e^{-\gamma}\exp(\mathcal{L}(\boldsymbol{Z}\to W)). \tag{67}$$

Substituting (67) into (56) and setting $\gamma = \mathcal{L}(\boldsymbol{Z}\to W) + \log(2/\delta)$, we conclude that with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$,

$$|\mathrm{gen}(W,\boldsymbol{Z})| \leq \sqrt{\frac{2\sigma^2}{n}\left(\mathcal{L}(\boldsymbol{Z}\to W) + \log 2 + 2\log\frac{2}{\delta}\right)}. \tag{68}$$

This coincides with the maximal leakage bound in (41) up to a $(2\sigma^2/n)\log 2$ term inside the square root, and with [21, Cor. 10] up to a $(2\sigma^2/n)\log(4/\delta)$ term inside the square root.

IV. Generalization Bounds for the Random-Subset Setting

We now consider the random-subset setting described in Section II. For this setting, we will require the stronger assumption that the loss function $\ell(\cdot,\cdot)$ is bounded, rather than the sub-Gaussian assumption in Section III. As detailed in the proof of Theorem 4 below, boundedness will be crucial to establish an inequality similar to (17) for the case in which the expectation over $\widetilde{\boldsymbol{Z}}$ is replaced by an expectation over the selection random variable $\boldsymbol{S}$.

The bounds in this section will depend on the conditional information density (2). Intuitively, rather than asking how much information on the training data $\boldsymbol{Z}$ can be inferred from W, we instead ask how much information W reveals about whether $\tilde{Z}_i$ or $\tilde{Z}_{i+n}$ has been used for training, for $i = 1,\ldots,n$, given the knowledge of $\widetilde{\boldsymbol{Z}}$. We will make this intuition more precise and highlight the advantages of the random-subset approach when we compare the generalization error bounds obtained in this section to the ones in Section III, under the assumption of a bounded loss function.

As in Section III, the generalization bounds in this section will take different forms: average generalization bounds, PAC-Bayesian bounds, and single-draw bounds. The average bound for the random-subset setting has a form similar to (14), namely

$$\left|\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}[\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))]\right| \leq \epsilon. \tag{69}$$

For the PAC-Bayesian and single-draw settings, it will turn out to be convenient to first obtain probabilistic bounds on the following quantity:

$$\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S}) = \frac{1}{n}\sum_{i=1}^{n}\big(\ell(W,Z_i(\bar{S}_i)) - \ell(W,Z_i(S_i))\big). \tag{70}$$

Here, $\bar{\boldsymbol{S}}$ is a vector whose entries are modulo-2 complements of the entries of $\boldsymbol{S}$. As a consequence, $\boldsymbol{Z}(\bar{\boldsymbol{S}})$ contains all the elements of $\widetilde{\boldsymbol{Z}}$ that are not in $\boldsymbol{Z}(\boldsymbol{S})$. So, instead of comparing the loss on the training data to the expected loss on a new sample, we compare it to a test loss, i.e., the loss on n samples that are independent of W. Note that quantities similar to (70) are what one computes when empirically assessing the generalization performance of a learning algorithm.

In the PAC-Bayesian setting, we will be interested in deriving bounds of the following form: with probability at least $1-\delta$ under $P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}} = P_{\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$,

$$\left|\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right]\right| \leq \epsilon. \tag{71}$$

Similarly, in the single-draw setting, the bounds of interest will be of the following form: with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}} = P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$,

$$\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \leq \epsilon. \tag{72}$$

As we establish in Theorem 3 below, the probabilistic bounds on $\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})$ given in (71) and (72) can be converted into probabilistic bounds on $\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))$ by adding a δ-dependent penalty term.

Theorem 3: Let $\widetilde{\boldsymbol{Z}} = (\tilde{Z}_1,\ldots,\tilde{Z}_{2n}) \in \mathcal{Z}^{2n}$ consist of $2n$ i.i.d. training samples generated from P_Z and let $\boldsymbol{S}$ be a random vector, independent of $\widetilde{\boldsymbol{Z}}$, with entries drawn independently from a $\mathrm{Bern}(1/2)$ distribution. Let $\boldsymbol{Z}(\boldsymbol{S})$ denote the subset of $\widetilde{\boldsymbol{Z}}$ obtained through $\boldsymbol{S}$ by the rule $Z_i(S_i) = \tilde{Z}_{i+S_i n}$, for $i = 1,\ldots,n$. Also, let $\bar{\boldsymbol{S}}$ be the modulo-2 complement of $\boldsymbol{S}$. Let $P_{W|\boldsymbol{Z}(\boldsymbol{S})}$ be a randomized learning algorithm.[3] Assume that $\ell(w,z)$ is bounded on $[a,b]$ for all $w \in \mathcal{W}$ and all $z \in \mathcal{Z}$. Also, assume that the

[3]Note that, by construction, W and $(\widetilde{\boldsymbol{Z}},\boldsymbol{S})$ are conditionally independent given $\boldsymbol{Z}(\boldsymbol{S})$.

following two probabilistic inequalities hold: with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \leq \epsilon_{\mathrm{SD}}(\delta) \tag{73}$$

and with probability at least $1-\delta$ under $P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\left|\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right]\right| \leq \epsilon_{\mathrm{PB}}(\delta). \tag{74}$$

Then, with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$|\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))| \leq \epsilon_{\mathrm{SD}}\left(\frac{\delta}{2}\right)+\sqrt{\frac{(b-a)^2}{2n}\log\frac{4}{\delta}} \tag{75}$$

and with probability at least $1-\delta$ under $P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\left|\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}[\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))]\right| \leq \epsilon_{\mathrm{PB}}\left(\frac{\delta}{2}\right)+\sqrt{\frac{(b-a)^2}{2n}\log\frac{4}{\delta}}. \tag{76}$$

Proof: Since $\ell(w,Z_i(S_i))$ is bounded on $[a,b]$ for all $i=1,\ldots,n$, it is $(b-a)/2$-sub-Gaussian for all $w\in\mathcal{W}$. From this, it follows that $L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(w)$ is $(b-a)/2\sqrt{n}$-sub-Gaussian for all $w\in\mathcal{W}$. Hence, using Hoeffding's inequality, stated in Lemma 2, we have that, for all $\epsilon>0$,

$$\left|L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)-\mathbb{E}_{P_Z}[\ell(W,Z)]\right| = \tag{77}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n}\ell\big(W,Z_i(\bar{S}_i)\big)-\mathbb{E}_{P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\big(W,Z_i(\bar{S}_i)\big)\right]\right| \geq \epsilon \tag{78}$$

with probability no larger than $\delta = 2\exp(-2\epsilon^2 n/(b-a)^2)$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$. From this it follows that, with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\left|L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)-\mathbb{E}_{P_Z}[\ell(W,Z)]\right| \leq \sqrt{\frac{(b-a)^2}{2n}\log\frac{2}{\delta}}. \tag{79}$$

Now note that, by the triangle inequality,

$$\begin{aligned}|\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))| \leq{}& \left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \\ &+\left|L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)-\mathbb{E}_{P_Z}[\ell(W,Z)]\right|.\end{aligned} \tag{80}$$

The result in (75) now follows by combining (73) and (79) via the union bound and performing the substitution $\delta\to\delta/2$. The proof of (76) follows along the same lines. ∎

We now turn to proving an exponential inequality similar to Theorem 1, but for the random-subset setting. This inequality will later be used to derive generalization bounds of the forms given in (69), (71), and (72).

Theorem 4: Consider the setting of Theorem 3. Then, for all $\lambda\in\mathbb{R}$,

$$\begin{aligned}\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg[\exp\Bigg(&\lambda\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S}) \\ &-\frac{\lambda^2(b-a)^2}{2n}-\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\Bigg)\Bigg]\leq 1.\end{aligned} \tag{81}$$

Proof: Due to the boundedness of $\ell(\cdot,\cdot)$, the random variable $\ell(W,Z_i(\bar{S}_i))-\ell(W,Z_i(S_i))$ is bounded on $[(a-b),(b-a)]$ for $i=1,\ldots,n$. As remarked in Lemma 2, this implies that it is $(b-a)$-sub-Gaussian, and that $\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})$ therefore is $(b-a)/\sqrt{n}$-sub-Gaussian. Furthermore, $\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})$ enjoys the symmetry property $\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})=-\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\bar{\boldsymbol{S}})$. From this, it follows that $\mathbb{E}_{P_{\boldsymbol{S}}}\left[\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right]=0$. By the definition of sub-Gaussianity, we therefore have that

$$\mathbb{E}_{P_{\boldsymbol{S}}}\left[\exp(\lambda\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S}))\right]\leq e^{\lambda^2(b-a)^2/2n}. \tag{82}$$

Reorganizing terms and taking the expectation with respect to $P_{W\widetilde{\boldsymbol{Z}}}$, we obtain

$$\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\left[\exp\left(\lambda\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})-\frac{\lambda^2(b-a)^2}{2n}\right)\right]\leq 1. \tag{83}$$

Now let $\mathcal{E}=\mathrm{supp}(P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}})$ be the support of $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$. Then, (83) implies that

$$\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\left[1_{\mathcal{E}}\exp\left(\lambda\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})-\frac{\lambda^2(b-a)^2}{2n}\right)\right]\leq 1. \tag{84}$$

Since $P_{W|\widetilde{\boldsymbol{Z}}}$ is induced from $P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ by the probability mass function $P_{\boldsymbol{S}}$, the probability distribution $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ is absolutely continuous with respect to $P_{W\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$. We can therefore perform a change of measure to $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$, as per [28, Prop. 17.1(4)], after which the desired result follows. ∎

Similar to the discussion in Remark 1, Theorem 4 holds *verbatim* with $P_{W|\widetilde{\boldsymbol{Z}}}$ replaced by an auxiliary conditional distribution $Q_{W|\widetilde{\boldsymbol{Z}}}$, provided that a suitable absolute continuity assumption holds. This is detailed in the following remark.

Remark 4: Consider the setting of Theorem 3. Also, assume that the absolute continuity assumption $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\ll Q_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$ holds for some conditional distribution $Q_{W|\widetilde{\boldsymbol{Z}}}$ on $\mathcal{W}$. Then,

$$\begin{aligned}\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg[\exp\Bigg(&\lambda\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})-\frac{\lambda^2(b-a)^2}{2n} \\ &-\log\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}Q_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\Bigg)\Bigg]\leq 1.\end{aligned} \tag{85}$$

The proof of (85) is exactly the same as the proof of Theorem 4, except that we choose $Q_{W|\widetilde{\boldsymbol{Z}}}$ in place of $P_{W|\widetilde{\boldsymbol{Z}}}$ in (83).

For the bounds that we will later derive, the optimal choice is $Q_{W|\widetilde{\boldsymbol{Z}}}=P_{W|\widetilde{\boldsymbol{Z}}}$. However, similar to the standard setting, this choice is not always feasible when one is interested in numerically evaluating the bounds. While it is technically possible to compute $P_{W|\widetilde{\boldsymbol{Z}}}$ for a given instance of $\widetilde{\boldsymbol{Z}}$ by marginalizing out $\boldsymbol{S}$, this would involve executing the probabilistic learning algorithm $P_{W|\boldsymbol{Z}(\boldsymbol{S})}$ a total of 2^n times. For many algorithms, this is prohibitively expensive from a computational standpoint. Therefore, it can be convenient to have the choice of relaxing the bound by expressing it in terms of some auxiliary distribution $Q_{W|\widetilde{\boldsymbol{Z}}}$, suitably chosen so as to trade accuracy with computational complexity.

We also note that the assumption of bounded loss in Theorem 4 can be relaxed, and an exponential inequality can be derived for unbounded loss functions satisfying the conditions specified in the following remark.

Remark 5: Assume that there exists a function $\Delta:\mathcal{Z}^2\to\mathbb{R}$ such that, for all $z_1,z_2\in\mathcal{Z}$ and all $w\in\mathcal{W}$, we have

$|\ell(w,z_1)-\ell(w,z_2)| \le \Delta(z_1,z_2)$. Let Z_1 and Z_2 be independent and distributed according to P_Z. Then, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg[\exp\Bigg(\lambda\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S}) - \frac{\lambda^2\,\mathbb{E}_{P_{Z_1Z_2}}\left[\Delta(Z_1,Z_2)^2\right]}{2n} - \imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\Bigg)\Bigg] \le 1. \quad (86)$$

The proof of (86) involves adapting the derivation in [14, p. 29] and using a change of measure argument. All the bounds presented in the remainder of this section for the bounded loss setting admit a counterpart for the unbounded loss setting, obtained by replacing $(b-a)^2$ with $\mathbb{E}_{P_{Z_1Z_2}}\left[\Delta(Z_1,Z_2)^2\right]$. Our choice to focus on the case of bounded loss functions in the remainder of this paper is justified by the fact that boundedness of $\mathbb{E}_{P_{Z_1Z_2}}\left[\Delta(Z_1,Z_2)^2\right]$ can be proven only for very specific cases (see [14, Sec. 5.4–5.6]).

In the remainder of this section, we will use Theorem 4 to derive an average generalization bound, as well as PAC-Bayesian bounds and single-draw bounds. We start with the average generalization bound.

A. Average Generalization Error Bounds

In the same spirit as Corollary 1, the following bound on the average generalization error, which is explicit in the conditional mutual information $I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$, is directly derived from Theorem 4.

Corollary 5: Under the setting of Theorem 3,

$$\left|\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))]\right| \le \sqrt{\frac{2(b-a)^2}{n} I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}. \quad (87)$$

Proof: Starting from (81), we apply Jensen's inequality, which results in

$$\exp\Bigg(\lambda\,\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right] - \frac{\lambda^2(b-a)^2}{2n} - \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\right]\Bigg) \le 1. \quad (88)$$

From (70), since W and $\boldsymbol{Z}(\bar{\boldsymbol{S}})$ are independent, it follows that $\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right] = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))]$. Furthermore, we have that

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\right] = D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}} \,|\, P_{\tilde{\boldsymbol{Z}}}) \quad (89)$$

$$= I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}}). \quad (90)$$

We therefore get, after reorganizing terms,

$$\frac{\lambda^2(b-a)^2}{2n} - \lambda\,\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))] + D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}} \,|\, P_{\tilde{\boldsymbol{Z}}}) \ge 0. \quad (91)$$

Since the parabola in λ described by (91) is nonnegative, its discriminant is nonpositive. This leads to

$$\mathbb{E}^2_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))] - \frac{2(b-a)^2}{n} D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}} \,|\, P_{\tilde{\boldsymbol{Z}}}) \le 0 \quad (92)$$

from which (87) follows directly. ∎

The bound in (87) recovers the result from [14, Cor. 5.2]. As detailed in Remark 4, we can substitute $Q_{W|\tilde{\boldsymbol{Z}}}$ for $P_{W|\tilde{\boldsymbol{Z}}}$ in (92) to obtain a more general but weaker bound in terms of the conditional relative entropy $D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}} \,||\, Q_{W|\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}} \,|\, P_{\tilde{\boldsymbol{Z}}})$, provided that an appropriate absolute continuity assumption is satisfied.

Under some conditions, the bound in Corollary 5 can be shown to be tighter than that in Corollary 1 for the case of a bounded loss function. Indeed, using the chain rule for mutual information, the Markov property $(\widetilde{\boldsymbol{Z}},\boldsymbol{S})$—$\boldsymbol{Z}(\boldsymbol{S})$—$W$, and the fact that $\boldsymbol{Z}(\boldsymbol{S})$ is a deterministic function of $(\widetilde{\boldsymbol{Z}},\boldsymbol{S})$, we can rewrite the bound in (22) as

$$|\mathrm{gen}(W,\boldsymbol{Z}(\boldsymbol{S}))| \le \sqrt{\frac{(b-a)^2}{2n} I(W;\boldsymbol{Z}(\boldsymbol{S}))} \quad (93)$$

$$= \sqrt{\frac{(b-a)^2}{2n}\left(I(W;\widetilde{\boldsymbol{Z}}) + I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\right)}. \quad (94)$$

Hence, if $I(W;\widetilde{\boldsymbol{Z}}) > 3I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$, the bound in Corollary 5 is tighter than that in Corollary 1. In particular, note that there are many practical scenarios in which the bound in Corollary 1 is vacuous because $I(W;\boldsymbol{Z}(\boldsymbol{S})) = \infty$. On the contrary, $I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \le n\log 2$.

B. PAC-Bayesian Generalization Error Bounds

We now turn to PAC-Bayesian bounds of the form given in (71). The next corollary provides bounds that are analogous to those in Corollary 2, but for the random-subset setting. The bounds in the corollary are novel, and extend known PAC-Bayesian bounds to the random-subset setting.

Corollary 6: Under the setting of Theorem 3, the following holds with probability at least $1-\delta$ under $P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ for all $t>0$:

$$\left|\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right]\right| \le \sqrt{\frac{2(b-a)^2}{n}\left(D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}}) + \log\frac{1}{\delta}\right)} \quad (95)$$

$$\le \sqrt{\frac{2(b-a)^2}{n}\left(\frac{\mathbb{E}^{1/t}_{P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}})^t\right]}{(\delta/2)^{1/t}} + \log\frac{2}{\delta}\right)}. \quad (96)$$

Here, the first inequality is a data-dependent bound, while the second provides a data-independent relaxation.

Proof: Since the proof follows along the same lines as that of Corollary 2, we only highlight the differences. We start from (81), apply Jensen's inequality with respect to $P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}$, and note that

$$\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\right] = D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}}). \quad (97)$$

To obtain (95), we use (30) and the same discriminant argument that was used after (32). To prove (96), we apply Markov's inequality to $D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}})^t$, similarly to (33). Combining the resulting inequality with (95) through the union bound and then performing the substitution $\delta \to \delta/2$, we obtain the desired result. ∎

For the case in which we set $t = 1$ in (96), we obtain $\mathbb{E}_{P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}})\right] = I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$. The corresponding bound extends the results in [14] by providing a PAC-Bayesian generalization error bound in terms of the conditional

mutual information $I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$. Similar to the discussion following Corollary 5, this bound is, under some conditions, tighter than the corresponding bounds for the standard setting in Corollary 2. Much like the moment bounds in (27) and (35), the bound in (96) illustrates a trade-off between the confidence and the tightness of the generalization estimate, mediated by the magnitude of the higher moments of $D(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W|\widetilde{\boldsymbol{Z}}})$. Also, as indicated in Remark 4, if the appropriate absolute continuity criterion is satisfied, we can replace $P_{W|\widetilde{\boldsymbol{Z}}}$ with $Q_{W|\widetilde{\boldsymbol{Z}}}$ in (95) and (96) to obtain more general bounds that are better suited for numerical evaluations.

C. Single-Draw Generalization Error Bounds

In this section, we will derive several bounds on the single-draw generalization error (72) in the random-subset setting. Three different approaches will be used to obtain these bounds. The first one relies on the exponential inequality given in Theorem 4, and results in a data-dependent bound from which several data-independent relaxations follow. The second one relies on Lemma 1, and allows us to derive a bound that is explicit in the tail of the conditional information density, similar to Theorem 2. Essentially equivalent versions of the data-independent relaxations obtainable via the first approach can be derived from this tail-based bound. The third approach, which is inspired by [21], builds on repeated applications of Hölder's inequality. This results in a family of data-independent bounds. Through this approach, we extend many of the results for bounded loss functions in [21] to the random-subset setting.

1) Generalization Bounds from the Exponential Inequality: In the next two corollaries, we derive novel bounds that are analogous to the ones in Corollaries 3 and 4, but for the random-subset setting.

Corollary 7: Under the setting of Theorem 3, the following holds with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ for all $t>0$:[4]

$$\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \leq \sqrt{\frac{2(b-a)^2}{n}\left(\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})+\log\frac{1}{\delta}\right)} \tag{98}$$

$$\leq \sqrt{\frac{2(b-a)^2}{n}\left(I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})+\frac{\widetilde{M}_t(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}{(\delta/2)^{1/t}}+\log\frac{2}{\delta}\right)}. \tag{99}$$

Here, the first inequality provides a data-dependent bound and the second is a data-independent relaxation. In (99), the term $\widetilde{M}_t(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$ is the tth root of the tth central moment of $\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$:

$$\widetilde{M}_t(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) = \mathbb{E}^{1/t}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\left|\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})-I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\right|^t\right]. \tag{100}$$

Proof: The proof is analogous to that of Corollary 3. We start by applying Markov's inequality in the form of (30) to (81), which, combined with the discriminant argument that was used after (37), results in (98). We then apply (38) with $U = \imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$. Combining the resulting inequality with (98) through the union bound, we obtain (99) after performing the substitution $\delta \to \delta/2$. ∎

[4]Note that the argument of the square root in (98) can be negative, but that this happens with probability at most δ. Therefore, the right-hand side of (98) is well-defined with probability at least $1-\delta$.

By increasing t in (99), a more benign δ-dependence can be obtained by letting the bound depend on higher central moments of $\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$. The tightness of the resulting bound depends on how well these higher moments are controlled. As usual, we can get more general bounds by replacing $P_{W|\widetilde{\boldsymbol{Z}}}$ with an arbitrary $Q_{W|\widetilde{\boldsymbol{Z}}}$, provided that a suitable absolute continuity assumption is satisfied.

Just as in Corollary 4, we can derive alternative data-independent relaxations for the data-dependent bound in (98). We present these novel bounds in the following corollary. The first bound is given in terms of $\mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}})$, the conditional maximal leakage (9). The second bound depends on the conditional Rényi divergence (4).

Corollary 8: Under the setting of Theorem 3, the following inequalities hold with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$:

$$\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \leq \sqrt{\frac{2(b-a)^2}{n}\left(\mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}})+2\log\frac{2}{\delta}\right)} \tag{101}$$

and, for all $\alpha,\gamma>1$ such that $1/\alpha+1/\gamma=1$,

$$\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \leq \Bigg[\frac{2(b-a)^2}{n}\times \left(2\log\frac{2}{\delta}+\frac{\alpha-1}{\alpha}D_\alpha(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}}\,||\,P_{W|\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}\,|\,P_{\widetilde{\boldsymbol{Z}}}) + \frac{\gamma-1}{\gamma}D_\gamma(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}}\,||\,P_{W|\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}\,|\,P_{\widetilde{\boldsymbol{Z}}})\right)\Bigg]^{1/2}. \tag{102}$$

Proof: Analogously to the proof of Corollary 4, we start from the inequality in (98) and bound $\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$. Markov's inequality implies that, with probability $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) = \log\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right]+\log\frac{1}{\delta}. \tag{103}$$

Replacing expectations with essential suprema, we get the upper bound

$$\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] \leq \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}}\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}|W\widetilde{\boldsymbol{Z}}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] \tag{104}$$

$$\leq \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}}\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right]. \tag{105}$$

Here, the second inequality holds due to the absolute continuity property $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}} \ll P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$. Using the union bound to combine (98) with the probabilistic inequality on $\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$ resulting from (103)–(105), we obtain (101) after performing the substitution $\delta\to\delta/2$.

To prove (102), we apply Markov's inequality and then perform a change of measure from $P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ to $P_{W|\widetilde{\boldsymbol{Z}}}$ to conclude

that, with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \le \log \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] + \log\frac{1}{\delta} \tag{106}$$

$$= \log \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\left(\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right)^2\right] + \log\frac{1}{\delta}. \tag{107}$$

Next, we apply Hölder's inequality thrice as follows. Let $\alpha,\gamma,\alpha',\gamma',\tilde{\alpha},\tilde{\gamma} > 1$ be constants such that $1/\alpha + 1/\gamma = 1/\alpha' + 1/\gamma' = 1/\tilde{\alpha} + 1/\tilde{\gamma} = 1$. Then,

$$\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\left(\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right)^2\right]$$

$$= \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\left[\exp(2\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}))\right] \tag{108}$$

$$\le \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\cdot\mathbb{E}^{1/\gamma}_{P_{\boldsymbol{S}}}\left[e^{\gamma\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right] \tag{109}$$

$$\le \mathbb{E}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\tilde{\alpha}}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\alpha}/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\times \right. \tag{110}$$

$$\left.\mathbb{E}^{1/\tilde{\gamma}}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\gamma}/\gamma}_{P_{\boldsymbol{S}}}\left[e^{\gamma\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\right]$$

$$\le \mathbb{E}^{1/\alpha'}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha'/\tilde{\alpha}}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\alpha}/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\right]\times \tag{111}$$

$$\mathbb{E}^{1/\gamma'}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\gamma'/\tilde{\gamma}}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\gamma}/\gamma}_{P_{\boldsymbol{S}}}\left[e^{\gamma\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\right].$$

We now substitute (111) into (107) and set $\alpha = \alpha' = \tilde{\alpha}$, which implies $\gamma = \gamma' = \tilde{\gamma}$. Using (4), we conclude that, with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \le \frac{\alpha-1}{\alpha}D_\alpha(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}}\,||\,P_{W|\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}\,|\,P_{\widetilde{\boldsymbol{Z}}}) + \frac{\gamma-1}{\gamma}D_\gamma(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}}\,||\,P_{W|\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}\,|\,P_{\widetilde{\boldsymbol{Z}}}) + \log\frac{1}{\delta}. \tag{112}$$

Combining (112) with (98) through the union bound and performing the substitution $\delta \to \delta/2$, we obtain (102). ∎

As usual, we can replace $P_{W|\widetilde{\boldsymbol{Z}}}$ by some auxiliary $Q_{W|\widetilde{\boldsymbol{Z}}}$ to get more general bounds, provided that a suitable absolute continuity assumption is satisfied.

The conditional maximal leakage bound in (101) can be tighter than the maximal leakage bound in [21, Cor. 9].[5] This is the case since the conditional maximal leakage $\mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}})$ is upper-bounded by the maximal leakage $\mathcal{L}(\boldsymbol{Z}(\boldsymbol{S})\to W)$. We prove this result in the following theorem.

Theorem 5: Consider the setting of Theorem 3. Then,

$$\mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}}) \le \mathcal{L}(\boldsymbol{Z}(\boldsymbol{S})\to W). \tag{113}$$

Proof: Because of the Markov property $(\widetilde{\boldsymbol{Z}},\boldsymbol{S})$—$\boldsymbol{Z}(\boldsymbol{S})$—$W$ and the fact that $\boldsymbol{Z}(\boldsymbol{S})$ is a deterministic function of $(\widetilde{\boldsymbol{Z}},\boldsymbol{S})$, the equality $\mathcal{L}(\boldsymbol{Z}(\boldsymbol{S})\to W) = \mathcal{L}((\widetilde{\boldsymbol{Z}},\boldsymbol{S})\to W)$ holds [26, Lem. 1]. We begin by moving one essential supremum outside of the expectation:

$$\mathcal{L}((\widetilde{\boldsymbol{Z}},\boldsymbol{S})\to W) = \log \mathbb{E}_{P_W}\left[\operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_W P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] \tag{114}$$

$$\ge \log \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}}\mathbb{E}_{P_W}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_W P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right]. \tag{115}$$

Now, let $\mathcal{E}_{\widetilde{\boldsymbol{Z}}} = \operatorname{supp}(P_{W|\widetilde{\boldsymbol{Z}}})$. It follows from (114) that

$$\mathcal{L}((\widetilde{\boldsymbol{Z}},\boldsymbol{S})\to W) \ge \log \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}}\mathbb{E}_{P_W}\left[1_{\mathcal{E}_{\widetilde{\boldsymbol{Z}}}}\operatorname*{ess\,sup}_{P_{\boldsymbol{S}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_W P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right]. \tag{116}$$

Next, we perform a change of measure from P_W to $P_{W|\widetilde{\boldsymbol{Z}}}$:

$$\log \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}}\mathbb{E}_{P_W}\left[1_{\mathcal{E}_{\widetilde{\boldsymbol{Z}}}}\operatorname*{ess\,sup}_{P_{\boldsymbol{S}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_W P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] = \log \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}}\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\frac{\mathrm{d}P_W}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}}\operatorname*{ess\,sup}_{P_{\boldsymbol{S}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_W P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] \tag{117}$$

Finally, since $\mathrm{d}P_W/\,\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}$ is independent of $\boldsymbol{S}$,

$$\log \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}}\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\frac{\mathrm{d}P_W}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}}\operatorname*{ess\,sup}_{P_{\boldsymbol{S}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_W P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] = \log \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}}\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}}}\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] \tag{118}$$

$$= \mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}}). \tag{119}$$

∎

2) Generalization Bounds from the Strong Converse: In this section, we will use Lemma 1 to derive single-draw generalization error bounds in the random-subset setting. In Theorem 6 below, we use Lemma 1 to obtain a novel bound in terms of the tail of the conditional information density $\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$.

Theorem 6: Under the setting of Theorem 3, with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\left|\widehat{\operatorname{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \le \left[\frac{2(b-a)^2}{n}\times \left(\gamma + \log\left(\frac{2}{\delta - P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\ge\gamma\right]}\right)\right)\right]^{1/2}. \tag{120}$$

This is valid for all γ such that the right-hand side is defined and real.

Proof: We will use Lemma 1 with $P = P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$, $Q = P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ and

$$\mathcal{E} = \{W,\widetilde{\boldsymbol{Z}},\boldsymbol{S} : \left|\widehat{\operatorname{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| > \epsilon\}. \tag{121}$$

Let the set $\mathcal{E}_{W\widetilde{\boldsymbol{Z}}} = \{\boldsymbol{S} : (W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\in\mathcal{E}\}$ denote the fibers of $\mathcal{E}$ with respect to W and $\widetilde{\boldsymbol{Z}}$. As noted in the proof of Theorem 4, $\widehat{\operatorname{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})$ is a $(b-a)/\sqrt{n}$-sub-Gaussian random variable with $\mathbb{E}_{P_{\boldsymbol{S}}}\left[\widehat{\operatorname{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right] = 0$. By using Hoeffding's

[5] Note that (101) provides a bound on $\widehat{\operatorname{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})$, whereas the bound in [21, Cor. 9] is on $\operatorname{gen}(W,\boldsymbol{Z})$. To compare the two, one therefore has to add the δ-dependent penalty term in Theorem 3.

inequality (Lemma 2), we therefore conclude that, for all W and $\widetilde{\boldsymbol{Z}}$,

$$P_{\boldsymbol{S}}\big[\mathcal{E}_{W\widetilde{\boldsymbol{Z}}}\big] \leq 2\exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right). \tag{122}$$

It follows that $Q[\mathcal{E}] \leq 2\exp\big(-n\epsilon^2/2(b-a)^2\big)$. Inserting this inequality into (10), we get

$$P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\left[\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| > \epsilon\right] \leq P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \geq \gamma\right] + 2\exp\left(\gamma - \frac{n\epsilon^2}{2(b-a)^2}\right). \tag{123}$$

We obtain the desired result by requiring the right-hand side of (123) to equal δ and solving for ϵ. ∎

Similar to the discussion in Remark 4, a completely analogous result holds with an auxiliary distribution $Q_{W|\widetilde{\boldsymbol{Z}}}$ in place of $P_{W|\widetilde{\boldsymbol{Z}}}$, provided that a suitable absolute continuity assumption is satisfied.

As for the bound in Theorem 2, the bound in (120) illustrates that the faster the rate of decay of the tail of the conditional information density, the sharper the generalization bound. Specifically, the parameter γ has to be chosen large enough so that the argument of the logarithm is positive, but a greater γ also contributes to an increased value for the bound.

The bound in Theorem 6 can be relaxed to give essentially equivalent versions of some of the previously presented data-independent bounds. We show this in the following remarks.

Remark 6 (Alternative derivation of the moment bound (99)*):* Markov's inequality implies that

$$P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \geq \gamma\right] \leq \frac{(\widetilde{M}_t(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}}))^t}{\left(\gamma - I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\right)^t} \tag{124}$$

where $\widetilde{M}_t(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})$ is defined in (100). Next, we set

$$\gamma = I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) + \frac{\widetilde{M}_t(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}{(\delta/2)^{1/t}} \tag{125}$$

which, once it is substituted into (124), implies the inequality $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \geq \gamma\right] \leq \delta/2$. Using this inequality in (120), we conclude that, with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \leq \left[\frac{2(b-a)^2}{n}\left(I(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) + \frac{\widetilde{M}_t(W;\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}{(\delta/2)^{1/t}} + \log\frac{4}{\delta}\right)\right]^{1/2}. \tag{126}$$

This coincides with the bound in (99), up to a $(2(b-a)^2/n)\log 2$ term inside the square root.

Remark 7 (Alternative derivation of the conditional maximal leakage bound (101)*):* Note that

$$P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \geq \gamma\right] \leq P_{W\widetilde{\boldsymbol{Z}}}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}|W\widetilde{\boldsymbol{Z}}}} e^{\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})} > e^{\gamma}\right] \tag{127}$$

$$\leq \operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}} P_{W|\widetilde{\boldsymbol{Z}}}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}|W\widetilde{\boldsymbol{Z}}}} e^{\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})} > e^{\gamma}\right]. \tag{128}$$

By upper-bounding the ess sup as in (105) and using Markov's inequality, we conclude that

$$P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \geq \gamma\right] \leq e^{\gamma}\operatorname*{ess\,sup}_{P_{\widetilde{\boldsymbol{Z}}}} \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}}} e^{\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right] \tag{129}$$

$$= \exp\Big(\mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}}) - \gamma\Big). \tag{130}$$

Setting $\gamma = \mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}}) + \log(2/\delta)$ and substituting the resulting upper-bound on the probability $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}\left[\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}}) \geq \gamma\right]$ into (120), we conclude that, with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \leq \left[\frac{2(b-a)^2}{n}\left(\mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}}) + \log 2 + 2\log\frac{2}{\delta}\right)\right]^{1/2}. \tag{131}$$

This recovers the conditional maximal leakage bound in (101), up to a $(2(b-a)^2/n)\log 2$ term inside the square root.

3) Generalization Bounds from a Hölder-Based Inequality: We now present a third approach to obtain data-independent single-draw bounds in the random-subset setting. The approach is based on a proof technique developed in [21], where similar bounds are derived in the standard setting. We first prove a useful inequality in Theorem 7, from which several generalization bounds follow.

Theorem 7: Under the setting of Theorem 3, for all constants $\alpha,\gamma,\alpha',\gamma',\tilde{\alpha},\tilde{\gamma} > 1$ such that $1/\alpha + 1/\gamma = 1/\alpha' + 1/\gamma' = 1/\tilde{\alpha} + 1/\tilde{\gamma} = 1$ and all measurable sets $\mathcal{E} \in \mathcal{W}\times\mathcal{Z}^{2n}\times\{0,1\}^n$,

$$P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}] \leq \mathbb{E}^{1/\tilde{\gamma}}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\gamma}/\gamma'}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[P^{\gamma'/\gamma}_{\boldsymbol{S}}\big[\mathcal{E}_{W\widetilde{\boldsymbol{Z}}}\big]\right]\right] \times \mathbb{E}^{1/\tilde{\alpha}}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\alpha}/\alpha'}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha'/\alpha}_{P_S}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\right]. \tag{132}$$

Here, $\mathcal{E}_{W\widetilde{\boldsymbol{Z}}} = \{\boldsymbol{S} : (W,\widetilde{\boldsymbol{Z}},\boldsymbol{S}) \in \mathcal{E}\}$ denotes the fibers of $\mathcal{E}$ with respect to W and $\widetilde{\boldsymbol{Z}}$.

Proof: First, we rewrite $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}]$ in terms of the expectation of the indicator function $1_{\mathcal{E}}$ and perform a change of measure:

$$P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}] = \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[1_{\mathcal{E}}\cdot\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] \tag{133}$$

$$= \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\left[1_{\mathcal{E}}\cdot e^{\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]. \tag{134}$$

To obtain the desired result, we apply Hölder's inequality thrice. Let $\alpha,\gamma,\alpha',\gamma',\tilde{\alpha},\tilde{\gamma} > 1$ be constants such that $1/\alpha + 1/\gamma = 1/\alpha' + 1/\gamma' = 1/\tilde{\alpha} + 1/\tilde{\gamma} = 1$. Then,

$$P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}] \leq \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\gamma}_{P_{\boldsymbol{S}}}\big[1_{\mathcal{E}_{W\widetilde{\boldsymbol{Z}}}}\big]\cdot\mathbb{E}^{1/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right] \tag{135}$$

$$\leq \mathbb{E}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\gamma'}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[P^{\gamma'/\gamma}_{\boldsymbol{S}}\big[\mathcal{E}_{W\widetilde{\boldsymbol{Z}}}\big]\right]\times \mathbb{E}^{1/\alpha'}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha'/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\right] \tag{136}$$

$$\leq \mathbb{E}^{1/\tilde{\gamma}}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\gamma}/\gamma'}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[P^{\gamma'/\gamma}_{\boldsymbol{S}}\big[\mathcal{E}_{W\widetilde{\boldsymbol{Z}}}\big]\right]\right]\times \mathbb{E}^{1/\tilde{\alpha}}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\alpha}/\alpha'}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha'/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\right]. \tag{137}$$

■

Similar to the discussion in Remark 4, the result in Theorem 7 would still hold if we were to substitute an auxiliary distribution $Q_{W|\widetilde{\boldsymbol{Z}}}$ for $P_{W|\widetilde{\boldsymbol{Z}}}$, provided that a suitable absolute continuity condition is satisfied.

By choosing particular values for the three free parameters in the inequality (132), we can derive generalization bounds in terms of various information-theoretic quantities. We will focus on a bound that depends on the conditional α-mutual information $I_\alpha(W;\boldsymbol{S}\,|\,\widetilde{\boldsymbol{Z}})$, which can be relaxed to obtain a bound in terms of the conditional Rényi divergence $D_\alpha(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}}\,||\,P_{W|\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}\,|\,P_{\widetilde{\boldsymbol{Z}}})$ or be specialized to obtain a bound that depends on the conditional maximal leakage $\mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}})$.

Corollary 9: Under the setting of Theorem 3, the following holds with probability at least $1-\delta$ under $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ for all $\alpha>1$:

$$\left|\widehat{\mathrm{gen}}(W,\widetilde{\boldsymbol{Z}},\boldsymbol{S})\right| \le \left[\frac{2(b-a)^2}{n}\left(I_\alpha(W;\boldsymbol{S}\,|\,\widetilde{\boldsymbol{Z}}) + \log 2 + \frac{\alpha}{\alpha-1}\log\frac{1}{\delta}\right)\right]^{1/2}. \quad (138)$$

Proof: In (132), set $\tilde{\alpha}=\alpha$ and let $\alpha'\to 1$, which implies that $\tilde{\gamma}=\gamma$ and $\gamma'\to\infty$. Also, let $\mathcal{E}$ be the error event (121). For this choice of parameters, the second factor in (132) reduces to

$$\mathbb{E}^{1/\alpha}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\alpha}_{P_S}\left[\exp\Big(\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})\Big)\right]\right]\right] = \exp\left(\frac{\alpha-1}{\alpha}I_\alpha(W;\boldsymbol{S}\,|\,\widetilde{\boldsymbol{Z}})\right). \quad (139)$$

Furthermore, we can bound $P_{\boldsymbol{S}}\big[\mathcal{E}_{W\widetilde{\boldsymbol{Z}}}\big]$ in the first factor in (132) by using (122). Substituting (122) into the first factor in (132), we conclude that

$$\lim_{\gamma'\to\infty}\mathbb{E}^{1/\gamma}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\gamma/\gamma'}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[P^{\gamma'/\gamma}_{\boldsymbol{S}}\big[\mathcal{E}_{W\widetilde{\boldsymbol{Z}}}\big]\right]\right] = \mathbb{E}^{1/\gamma}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\left(\operatorname*{ess\,sup}_{P_{W|\widetilde{\boldsymbol{Z}}}} P^{1/\gamma}_{\boldsymbol{S}}\big[\mathcal{E}_{W\widetilde{\boldsymbol{Z}}}\big]\right)^{\gamma}\right] \quad (140)$$

$$\le \left(2\exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right)\right)^{1/\gamma}. \quad (141)$$

By substituting (139) and (140) into (132), noting that $1/\gamma = (\alpha-1)/\alpha$, we conclude that

$$P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}] \le \left(2\exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right)\right)^{\frac{\alpha-1}{\alpha}} \times \exp\left(\frac{\alpha-1}{\alpha}I_\alpha(W;\boldsymbol{S}\,|\,\widetilde{\boldsymbol{Z}})\right). \quad (142)$$

We obtain the desired result by requiring the right-hand side of (142) to equal δ and solving for ϵ. ■

As usual, we can obtain a more general version of Corollary 9 by replacing $P_{W|\widetilde{\boldsymbol{Z}}}$ with an auxiliary distribution $Q_{W|\widetilde{\boldsymbol{Z}}}$, provided that a suitable absolute continuity assumption is satisfied.

We can also obtain a bound in terms of the conditional maximal leakage by letting $\alpha\to\infty$ in (138) and using that $\lim_{\alpha\to\infty} I_\alpha(W;\boldsymbol{S}\,|\,\widetilde{\boldsymbol{Z}}) = \mathcal{L}(\boldsymbol{S}\to W|\widetilde{\boldsymbol{Z}})$. The resulting bound is tighter than the conditional maximal leakage bound obtained in (101) by a $(2(b-a)^2/n)\log(2/\delta)$ term inside the square root.

Furthermore, the conditional α-mutual information that appears in (138) can be relaxed to obtain a novel bound in terms of the conditional Rényi divergence of order α. Indeed, by Jensen's inequality, the following holds for $\alpha>1$:

$$I_\alpha(W;\boldsymbol{S}\,|\,\widetilde{\boldsymbol{Z}}) = \frac{1}{\alpha-1}\log\mathbb{E}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\right] \quad (143)$$

$$\le \frac{1}{\alpha-1}\log\mathbb{E}_{P_{\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}}}\left[\mathbb{E}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\widetilde{\boldsymbol{Z}})}\right]\right]\right] \quad (144)$$

$$= D_\alpha(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}}\,||\,P_{W|\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}\,|\,P_{\widetilde{\boldsymbol{Z}}}). \quad (145)$$

The conditional Rényi divergence bound obtained by substituting (145) into (138) is different from the one in (102), and there is no clear ordering between them in general. The two bounds can, however, be directly compared if we set $\alpha=\gamma=2$, or if we let $\alpha\to\infty$, and hence $\gamma\to 1$. For both of these choices of parameters, the conditional Rényi divergence bound obtained from (138) is tighter than (102) by a $(2(b-a)^2/n)\log(2/\delta)$ term inside the square root.

V. Conclusion

We have presented a general framework for deriving generalization bounds for probabilistic learning algorithms, not only in the average sense, but also for the PAC-Bayesian and the single-draw setup. Using this framework, we recovered several known results, and also presented new ones. Due to its unifying nature, the framework enables the transfer of methods for tightening bounds in one setup to the other two setups. In particular, by reobtaining previously known results, we showed that our framework subsumes proofs that are based on the Donsker-Varadhan variational formula for relative entropy [8, Thm. 1], [17, Prop. 3], on Hölder's inequality [21, Thm. 1], and on the data-processing inequality [16, Thm. 8], [21, p. 10]. We further demonstrated the versatility of the framework by applying it to the random-subset setting recently introduced by Steinke and Zakynthinou [14]. In doing so, we were able to extend the bounds on the average generalization error obtained in [14] to the PAC-Bayesian setup and the single-draw setup. In addition to this, we used tools inspired by binary hypothesis testing to derive generalization bounds in terms of the tail of the conditional information density. We also obtained novel bounds in terms of the conditional maximal leakage and the conditional α-mutual information by adapting a proof technique due to Esposito *et al.* [21] to the random-subset setting.

As pointed out throughout this paper, the numerical evaluation of the presented generalization bounds often requires one to replace the marginal distribution P_W (or $P_{W|\widetilde{\boldsymbol{Z}}}$ in the random-subset setting) with a suitably chosen auxiliary distribution that can be computed without *a priori* knowledge of the data distribution P_Z. Some possible choices, in the context of deep neural networks, are provided in [13], [18], [19], [30]. Specifically, in [30], we evaluate the PAC-Bayesian bound in (95) and the single-draw bound in (98) for neural networks trained on MNIST and Fashion-MNIST using stochastic gradient descent. The numerical experiments illustrate that the resulting bounds

are non-vacuous for the setups considered, and match the best bounds available in the literature [19]. While the results in [30] appear promising, they still do not provide much insight into how to design neural networks. Thus, the extent to which information-theoretic bounds such as the ones presented in this paper can guide the design of modern machine learning algorithms remains to be investigated.

REFERENCES

[1] F. Hellström and G. Durisi, "Generalization error bounds via mth central moments of the information density," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Los Angeles, CA, June 2020.

[2] D. A. McAllester, "Some PAC-Bayesian theorems," in *Proc. Conf. Learn. Theory (COLT)*, Madison, WI, July 1998, pp. 230–234.

[3] B. Guedj, "A primer on PAC-Bayesian learning," *arXiv*, Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.05353

[4] O. Catoni, *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. IMS Lecture Notes Monogr. Ser., 2007, vol. 56.

[5] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[6] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Toulon, France, Apr. 2017.

[7] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proc. Artif. Intell. Statist. (AISTATS)*, Cadiz, Spain, May 2016.

[8] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, 2017.

[9] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, July 2019.

[10] A. R. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, Canada, Dec. 2018.

[11] R. van Handel, *Probability in High Dimension*, 2016. [Online]. Available: http://web.math.princeton.edu/%7Ervan/APC550.pdf

[12] A. Achille and S. Soatto, "Emergence of Invariance and Disentanglement in Deep Representations," *J. of Mach. Learn. Res.*, vol. 19, pp. 1–34, Sep. 2018.

[13] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for SGLD via data-dependent estimates," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

[14] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," *arXiv*, Feb. 2020. [Online]. Available: https://arxiv.org/abs/2001.09122

[15] 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," *arXiv*, April 2020. [Online]. Available: http://arxiv.org/abs/2004.12983

[16] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," *J. of Mach. Learn. Res.*, vol. 83, pp. 25–55, Apr. 2018.

[17] B. Guedj and L. Pujol, "Still no free lunches: the price to pay for tighter PAC-Bayes bounds," *arXiv*, Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.04460

[18] G. K. Dziugaite and D. M. Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," in *Proc. Conf. Uncertainty in Artif. Intell. (UAI)*, Sydney, Australia, Aug. 2017.

[19] G. K. Dziugaite, K. Hsu, W. Gharbieh, and D. M. Roy, "On the role of data in PAC-Bayes bounds," *arXiv*, June 2020. [Online]. Available: https://arxiv.org/abs/2006.10929

[20] W. Zhou, V. Veitch, M. Austern, R. P. Adams, and P. Orbanz, "Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach," in *Proc. Int. Conf. Learn. Representations (ICLR)*, New Orleans, LA, May 2019.

[21] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rènyi f-divergences and maximal leakage," *arXiv*, Dec. 2019. [Online]. Available: http://arxiv.org/abs/1912.01439

[22] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, Aug. 2014.

[23] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, Canada, Dec. 2015.

[24] T. Van Erven and P. Harrëmos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[25] S. Verdú, "α-mutual information," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, Feb. 2015.

[26] I. Issa, S. Kamath, and A. B. Wagner, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[27] M. Tomamichel and M. Hayashi, "Operational interpretation of Rényi information measures via composite hypothesis testing against product and Markov distributions," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1064–1082, Feb. 2018.

[28] Y. Polyanskiy and Y. Wu, *Lecture Notes On Information Theory*, 2019. [Online]. Available: http://www.stat.yale.edu/%7Eyw562/teaching/itlectures.pdf

[29] M. J. Wainwright, *High-Dimensional Statistics: a Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[30] F. Hellström and G. Durisi, "Nonvacuous loss bounds with fast rates for neural networks via conditional information measures," *arXiv*, October 2020. [Online]. Available: https://arxiv.org/abs/2010.11552

Fredrik Hellström (S'20) received the B.Sc. and M.Sc. degrees in physics from the University of Gothenburg, Gothenburg, Sweden in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree in electrical engineering at Chalmers University of Technology, Gothenburg, Sweden. His research interests are in the areas of statistical learning theory and machine learning.

Giuseppe Durisi (S'02–M'06–SM'12) received the Laurea degree *summa cum laude* and the Doctor degree both from Politecnico di Torino, Italy, in 2001 and 2006, respectively. From 2002 to 2006, he was with Istituto Superiore Mario Boella, Torino, Italy. From 2006 to 2010 he was a postdoctoral researcher at ETH Zurich, Zurich, Switzerland. In 2010, he joined Chalmers University of Technology, Gothenburg, Sweden, where he is now full professor with the Communication Systems Group. He is co-director of Chalmers ICT Area of Advance, and of Chalmers AI Research Center. Dr. Durisi is a senior member of the IEEE. He is the recipient of the 2013 IEEE ComSoc Best Young Researcher Award for the Europe, Middle East, and Africa Region, and is co-author of a paper that won a "student paper award" at the 2012 International Symposium on Information Theory, and of a paper that won the 2013 IEEE Sweden VT-COM-IT joint chapter best student conference paper award. In 2015, he joined the editorial board of the IEEE Transactions on Communications as associate editor. From 2011 to 2014, he served as publications editor for the IEEE Transactions on Information Theory. His research interests are in the areas of communication and information theory and machine learning.