Generalization Error Bounds via mth Central Moments of the Information Density

Citation for the original published paper (version of record):
Hellström, F., Durisi, G. (2020). Generalization Error Bounds via mth Central Moments of the Information Density. IEEE International Symposium on Information Theory - Proceedings, 2020-June: 2741-2746. http://dx.doi.org/10.1109/ISIT44484.2020.9174475

Generalization Error Bounds via mth Central Moments of the Information Density

Fredrik Hellström, Giuseppe Durisi

Department of Electrical Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden

***Abstract*—We present a general approach to deriving bounds on the generalization error of randomized learning algorithms. Our approach can be used to obtain bounds on the average generalization error as well as bounds on its tail probabilities, both for the case in which a new hypothesis is randomly generated every time the algorithm is used—as often assumed in the probably approximately correct (PAC)-Bayesian literature—and in the single-draw case, where the hypothesis is extracted only once.**

For this last scenario, we present a novel bound that is explicit in the central moments of the information density. The bound reveals that the higher the order of the information density moment that can be controlled, the milder the dependence of the generalization bound on the desired confidence level.

Furthermore, we use tools from binary hypothesis testing to derive a second bound, which is explicit in the tail of the information density. This bound confirms that a fast decay of the tail of the information density yields a more favorable dependence of the generalization bound on the confidence level.

I. Introduction

A recent line of research, initiated by the work of Russo and Zou [1] and then followed by many recent contributions [2]–[5], has focused on obtaining bounds on the generalization error of randomized learning algorithms in terms of information-theoretic quantities, such as mutual information. The resulting bounds are *deterministic*, i.e., data-independent, and allow one to assess the speed of convergence of a given learning algorithm in terms of sample complexity [6, p. 44].

A parallel development has taken place in the machine learning and statistics community, where the probably approximately correct (PAC)-Bayesian framework, pioneered by McAllester [7], has resulted in several upper bounds on the generalization error. These bounds, which are expressed in terms of the relative entropy between a prior and a posterior distribution on the hypothesis class (see, e.g., [8] for a recent review), are typically *empirical*, i.e., data-dependent, and can be used to design learning algorithms [9].

One difficulty in comparing the bounds on the generalization error available in the literature is that they sometimes pertain to different quantities. To illustrate this point, we need to introduce some key quantities, which will be used in the remainder of the paper. Following the standard terminology in statistical learning theory, we 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. A training data set $Z^n = [Z_1, \dots, Z_n]$ is a set of n i.i.d. samples drawn from a distribution P_Z defined on $\mathcal{Z}$. We denote by P_{Z^n} the product distribution induced by P_Z. A randomized learning algorithm is characterized by a conditional probability distribution $P_{W|Z^n}$ on $\mathcal{W}$. Finally, we let the generalization error for a given hypothesis w be defined as the difference between the population and empirical risks

$$\text{gen}(w, z^n) = \frac{1}{n}\sum_{k=1}^{n} \ell(w, z_k) - \mathbb{E}_{P_Z}[\ell(w, Z)] . \tag{1}$$

Throughout the paper, we shall assume that the loss function $\ell(w, Z)$ is σ-subgaussian [10, Def. 2.2] under P_Z for all $w \in \mathcal{W}$.

The line of work initiated with [1] deals with bounding the average generalization error

$$\mathbb{E}_{P_{WZ^n}}[\text{gen}(W, Z^n)] . \tag{2}$$

Specifically, upper bounds on the absolute value of this quantity were first presented in [1] and then improved in [2, Thm. 1] and [4, Prop. 1].

On the contrary, the PAC-Bayesian approach seeks lower bounds on the probability [8]

$$P_{Z^n}\left[\left|\mathbb{E}_{P_{W|Z^n}}[\text{gen}(W, Z^n)]\right| \leq \epsilon\right] . \tag{3}$$

Characterizing such a probability, which is in the spirit of the PAC framework, is relevant when a new hypothesis W is drawn from $P_{W|Z^n}$ every time the algorithm is used. As can be verified by, e.g., comparing the proof of [2, Lemma 1] and the proof of [11, Prop. 3],[1] for the subgaussian case, one can obtain bounds both on (2) and on (3) that are explicit in the mutual information $I(W; Z^n)$ and in the relative entropy $D(P_{W|Z^n} \,||\, P_W)$, respectively, by using the Donsker-Varadhan variational formula for relative entropy.

One may also be interested in the scenario in which the hypothesis W is drawn from $P_{W|Z^n}$ only once, i.e., it is kept fixed for all uses of the algorithm. In such a scenario, which, following the terminology used in [9, p. 12], we shall refer to as a *single-draw* scenario, the probability of interest is

$$P_{WZ^n}[|\text{gen}(W, Z^n)| \leq \epsilon] . \tag{4}$$

Bounds on this probability that depend on the mutual information $I(W; Z^n)$ were provided in [2, Thm. 3] and [3]. Several novel bounds, which are explicit in information-theoretic quantities such as f-divergence, α-mutual information, and maximal leakage, were recently derived in [5]. Interestingly, all these bounds make use of a different set of tools compared with the ones used

[1] For the case in which the prior and posterior distributions in [11, Prop. 3] are set to P_W and $P_{W|Z^n}$, respectively.

to establish bounds on (2) and (3), with one of the main ingredients being the data processing inequality for f-divergences.

Furthermore, they yield drastically different estimates for the generalization error. Specifically, let us assume that we want (4) to be greater than $1-\delta$ where, throughout the paper, $\delta \in (0,1)$. Then a slight refinement of the analysis in [3] yields the following bound on ϵ:

$$\epsilon \geq \sqrt{\frac{2\sigma^2}{n}\left(\frac{I(W;Z^n)+H_b(\delta)}{\delta}+\log 2\right)}. \tag{5}$$

Here, $H_b(\delta)$ denotes the binary entropy function. Throughout the paper, $\log(\cdot)$ denotes the natural logarithm. In contrast, the analysis in [5, Cor. 5], yields the following bound for $\alpha > 1$:

$$\epsilon \geq \sqrt{\frac{2\sigma^2}{n}\left[I_\alpha(W;Z^n)+\log 2+\frac{\alpha}{\alpha-1}\log\frac{1}{\delta}\right]}. \tag{6}$$

Here, $I_\alpha(\cdot,\cdot)$ is the α-mutual information

$$I_\alpha(W;Z^n) = \frac{\alpha}{\alpha-1}\log \mathbb{E}_{P_{Z^n}}\left[\mathbb{E}^{1/\alpha}_{P_W}\left[\left(\frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}\right)^{\alpha}\right]\right], \tag{7}$$

where $\mathrm{d}P_{WZ^n}/\,\mathrm{d}P_W P_{Z^n}$ is the Radon-Nikodym derivative. Note that, since $\lim_{\delta\to 0} H_b(\delta)/\delta + \log\delta = 1$, the dependence of ϵ on δ in (5) is of order $1/\sqrt{\delta}$. In contrast, it is of order $\sqrt{(\alpha/(\alpha-1))\log(1/\delta)}$ in (6), which is typically more favorable. For example, in the limit $\alpha \to \infty$, the α-mutual information converges to the maximal leakage [12, Thm. 1], and ϵ depends on δ only through the term $\sqrt{\log(1/\delta)}$.

The analysis in [5], however, does not reveal why using α-mutual information rather than mutual information results in a more benign dependence of the generalization error on the confidence parameter δ. Moreover, the choice $\alpha = 1$, for which $I_\alpha(W;Z^n)$ reduces to $I(W;Z^n)$, renders the bound in (6) vacuous.

Contributions: Inspired by the treatment of the generalization error for the case of the $0-1$ loss function reported in [9], we present a single framework for deriving bounds on the generalization error that can be applied to both average and tail analyses, both of a PAC-Bayesian and single-draw flavor. As a product of our analysis, we obtain a probabilistic generalization error bound for the single-draw scenario, which results in the following bound on ϵ to guarantee that (4) is greater than $1-\delta$:

$$\epsilon \geq \sqrt{\frac{2\sigma^2}{n}\left(I(W;Z^n)+\frac{M_m(W;Z^n)}{(\delta/2)^{1/m}}+\log\frac{2}{\delta}\right)}. \tag{8}$$

Here,

$$M_m(W;Z^n) = \mathbb{E}^{1/m}_{P_{WZ^n}}\left[|\imath(W,Z^n)-I(W;Z^n)|^m\right] \tag{9}$$

is the mth root of the mth central moment of the information density

$$\imath(w,z^n) = \log\frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}(w,z^n). \tag{10}$$

The bound in (8) is derived as a data-independent relaxation of an underlying data-dependent bound. Comparing (5) with (8), we see that the existence of higher central moments of $\imath(W,Z^n)$ results in a more favorable scaling of the error bound with δ. This implies that one can obtain generalization error bounds that are explicit in the mutual information and have a more favorable dependence on δ than the one given in (5). In the limit $m\to\infty$, the dependence is of order $\sqrt{\log(1/\delta)}$, but the resulting bound is less tight than the maximal leakage bound in [5, Cor. 5]. However, through a more refined analysis, we recover the maximal leakage bound, up to a logarithmic term.

To shed further light on the role of the tail of the information density in determining the dependence of ϵ on δ, we derive an additional probabilistic single-draw bound, based on a change of measure argument [13, Thm. 12.5] that is used to establish strong converse bounds in binary hypothesis testing. It results in the following bound on ϵ:

$$\epsilon \geq \sqrt{\frac{2\sigma^2}{n}\left(\gamma+\log\left(\frac{2}{\delta-P_{WZ^n}[\imath(W,Z^n)\geq\gamma]}\right)\right)}. \tag{11}$$

Similar to (8), this bound reveals that for a fixed δ, low values of ϵ require fast-decaying tails of the information density random variable. Indeed, γ in (11) should be chosen sufficiently large to make the argument of the $\log$ positive. However, large values of γ also contribute to a large ϵ.

II. Bounds via a Subgaussian Inequality

In this section, we derive several types of bounds on the absolute value of the generalization error of a randomized learning algorithm. The following theorem gives an inequality that will later be used to derive both average and tail bounds for the generalization error.

Theorem 1: Let Z^n be i.i.d. according to P_Z. Assume that $\ell(w,Z)$ is σ-subgaussian under P_Z for all $w\in\mathcal{W}$. Assume that P_{WZ^n} is absolutely continuous with respect to $P_W P_{Z^n}$. Then, for all $\lambda\in\mathbb{R}$,

$$\mathbb{E}_{P_{WZ^n}}\left[\exp\left(\lambda\mathrm{gen}(W,Z^n)-\frac{\lambda^2\sigma^2}{2n}-\imath(W,Z^n)\right)\right]\leq 1. \tag{12}$$

Proof: Since $\ell(w,Z)$ is σ-subgaussian and the Z_i are i.i.d., the random variable $\frac{1}{n}\sum_{i=1}^n \ell(w,Z_i)$ is $\sigma/\sqrt{n}$-subgaussian, i.e.,

$$\mathbb{E}_{P_{Z^n}}\left[\exp\left(\lambda\left(\frac{1}{n}\sum_{i=1}^n\ell(w,Z_i)-\mathbb{E}_{P_Z}[\ell(w,Z)]\right)\right)\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2n}\right). \tag{13}$$

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

$$\mathbb{E}_{P_W P_{Z^n}}\left[\exp\left(\lambda\mathrm{gen}(W,Z^n)-\frac{\lambda^2\sigma^2}{2n}\right)\right]\leq 1. \tag{14}$$

Now, let E be the union of all sets $\mathcal{E}\in\mathcal{W}\times\mathcal{Z}^n$ such that $P_{WZ^n}(\mathcal{E})=0$, and let $\bar{E}$ denote its complement. It follows from (14) that

$$\mathbb{E}_{P_W P_{Z^n}}\left[1_{\bar{E}}\cdot\exp\left(\lambda\mathrm{gen}(W,Z^n)-\frac{\lambda^2\sigma^2}{2n}\right)\right]\leq 1, \tag{15}$$

where $1_{\bar{E}}$ is the indicator function of the set $\bar{E}$. To obtain (12), we perform a change of measure from $P_W P_{Z^n}$ to P_{WZ^n}, as per [13, Prop. 17.1(4)]. ■

We next show how the inequality (12) can be used to derive previously known and novel bounds on the generalization error.

A. Average Generalization Error

As a first corollary of Theorem 1, we derive a bound on the average generalization error (2), recovering the result in [2, Thm. 1].

Corollary 2: Under the assumptions of Theorem 1,

$$|\mathbb{E}_{P_{WZ^n}}[\mathrm{gen}(W,Z^n)]| \le \sqrt{\frac{2\sigma^2}{n} I(W;Z^n)}. \tag{16}$$

Proof: We apply Jensen's inequality to (12), which yields

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

Noting that $\mathbb{E}_{P_{WZ^n}}[\imath(W,Z^n)] = I(W;Z^n)$ we get, after taking the $\log$ of both sides of (17) and reorganizing terms, the nonnegative parabola in λ

$$\lambda^2\frac{\sigma^2}{2n} - \lambda\,\mathbb{E}_{P_{WZ^n}}[\mathrm{gen}(W,Z^n)] + I(W;Z^n) \ge 0. \tag{18}$$

Since the discriminant of a nonnegative parabola is nonpositive, we get

$$\mathbb{E}^2_{P_{WZ^n}}[\mathrm{gen}(W,Z^n)] - \frac{2\sigma^2}{n} I(W;Z^n) \le 0, \tag{19}$$

which yields the desired bound. ■

B. PAC-Bayesian Tail Bounds

Next, we use Theorem 1 to obtain two tail bounds on the absolute value of the generalization error averaged over $P_{W|Z^n}$ in (3). The first one, presented in Corollary 3, recovers a classical data-dependent PAC-Bayesian bound (see, e.g., [11, Prop. 3]) for the special case in which P_W is taken as the prior distribution and $P_{W|Z^n}$ is taken as the posterior distribution. The second one, presented in Corollary 4, is a relaxation of the first bound, which makes it data-independent. This bound, which depends on the mth moment of the relative entropy $D(P_{W|Z^n}\,||\,P_W)$, recovers the bound given in [3, App. A.3] for the case $m = 1$.

Corollary 3: Under the assumptions in Theorem 1, the following bound holds with probability at least $1-\delta$ under P_{Z^n}:

$$\begin{aligned}&\left|\mathbb{E}_{P_{W|Z^n}}\,\mathrm{gen}(W,Z^n)\right| \\ &\qquad \le \sqrt{\frac{2\sigma^2}{n}\left(D(P_{W|Z^n}\,||\,P_W) + \log\frac{1}{\delta}\right)}.\end{aligned} \tag{20}$$

Proof: Similarly to the proof of Corollary 2, we apply Jensen's inequality to (12), but now only with respect to the conditional expectation of W given Z^n. This yields

$$\begin{aligned}\mathbb{E}_{P_{Z^n}}\Bigg[\exp\bigg(&\lambda\,\mathbb{E}_{P_{W|Z^n}}[\mathrm{gen}(W,Z^n)] - \frac{\lambda^2\sigma^2}{2n} \\ &- D(P_{W|Z^n}\,||\,P_W)\bigg)\Bigg] \le 1,\end{aligned} \tag{21}$$

where we used that

$$\mathbb{E}_{P_{W|Z^n=z^n}}[i(W,z^n)] = D(P_{W|Z^n=z^n}\,||\,P_W). \tag{22}$$

Next, we use Markov's inequality in the following form: let $U \sim P_U$ be a nonnegative random variable s.t. $\mathbb{E}[U] \le 1$. Then

$$P_U[U > 1/\delta] < \mathbb{E}[U]\,\delta \le \delta. \tag{23}$$

Using (23) in (21), we conclude that

$$\begin{aligned}P_{Z^n}\Bigg[\exp\bigg(&\lambda\,\mathbb{E}_{P_{W|Z^n}}[\mathrm{gen}(W,Z^n)] - \frac{\lambda^2\sigma^2}{2n} \\ &- D(P_{W|Z^n}\,||\,P_W)\bigg) \le \frac{1}{\delta}\Bigg] \ge 1-\delta.\end{aligned} \tag{24}$$

Reorganizing terms, we obtain:

$$\begin{aligned}P_{Z^n}\Bigg[&\frac{\lambda^2\sigma^2}{2n} - \lambda\,\mathbb{E}_{P_{W|Z^n}}[\mathrm{gen}(W,Z^n)] \\ &+ D(P_{W|Z^n}\,||\,P_W) + \log\frac{1}{\delta} \ge 0\Bigg] \ge 1-\delta.\end{aligned} \tag{25}$$

The desired bound (20) now follows from the same discriminant analysis as in the proof of Corollary 2. ■

The bound in Corollary 3 is data-dependent because the upper bound on the generalization error depends on the specific instance of Z^n. In the next corollary, we apply Markov's inequality once more to make the bound data-independent.

Corollary 4: Under the assumptions in Theorem 1, the following bound holds with probability at least $1-\delta$ under P_{Z^n} for all $m > 0$:

$$\begin{aligned}&\left|\mathbb{E}_{P_{W|Z^n}}[\mathrm{gen}(W,Z^n)]\right| \\ &\le \sqrt{\frac{2\sigma^2}{n}\left(\frac{\mathbb{E}^{1/m}_{P_{Z^n}}\left[D(P_{W|Z^n}\,||\,P_W)^m\right]}{(\delta/2)^{1/m}} + \log\frac{2}{\delta}\right)}.\end{aligned} \tag{26}$$

Proof: Applying Markov's inequality to the random variable $D(P_{W|Z^n}\,||\,P_W)^m$, we obtain after some manipulations

$$\begin{aligned}P_{Z^n}\left[D(P_{W|Z^n}\,||\,P_W) \le \frac{\mathbb{E}^{1/m}_{P_{Z^n}}\left[D(P_{W|Z^n}\,||\,P_W)^m\right]}{\delta^{1/m}}\right] \\ \ge 1-\delta.\end{aligned} \tag{27}$$

We now observe that the two probability bounds (20) and (27) together with the union bound imply that, with probability at least $1-2\delta$ under P_{Z^n},

$$\begin{aligned}&\left|\mathbb{E}_{P_{W|Z^n}}[\mathrm{gen}(W,Z^n)]\right| \\ &\le \sqrt{\frac{2\sigma^2}{n}\left(\frac{\mathbb{E}^{1/m}_{P_{Z^n}}\left[D(P_{W|Z^n}\,||\,P_W)^m\right]}{\delta^{1/m}} + \log\frac{1}{\delta}\right)}.\end{aligned} \tag{28}$$

The desired result then follows by the substitution $\delta \to \delta/2$. ■

Note that when $m = 1$, we have

$$\mathbb{E}_{P_{Z^n}}\left[D(P_{W|Z^n}\,||\,P_W)\right] = I(W;Z^n) \tag{29}$$

and the bound (26) coincides with the one reported in [3, App. 3]. Some additional remarks on (26) are provided in Section II-D.

C. Single-Draw Probabilistic Bounds

We now use Theorem 1 to derive tail bounds on the absolute value of the single-draw generalization error in (4). As in Section II-B, we first state a data-dependent bound in Corollary 5. Then, we relax this to two different data-independent bounds in Corollaries 6 and 7. To the best of our knowledge, the first two bounds are novel, while the third recovers [5, Cor. 10] up to a logarithmic term.

Corollary 5: Under the assumptions in Theorem 1, the following bound holds with probability at least $1-\delta$ under P_{WZ^n}:[2]

$$|\text{gen}(W, Z^n)| \leq \sqrt{\frac{2\sigma^2}{n}\left(\imath(W, Z^n) + \log\frac{1}{\delta}\right)}. \tag{30}$$

Proof: Applying Markov's inequality (23) directly to (12), we conclude that

$$P_{WZ^n}\left[\exp\left(\lambda \text{gen}(W, Z^n) - \frac{\lambda^2\sigma^2}{2n} - \imath(W, Z^n)\right) \leq \frac{1}{\delta}\right] \geq 1-\delta, \tag{31}$$

from which the desired result follows by the same discriminant analysis as in the proof of Corollary 2. ■

Corollary 6: Under the assumptions in Theorem 1, the following bound holds with probability at least $1-\delta$ under P_{WZ^n}:

$$|\text{gen}(W, Z^n)| \leq \sqrt{\frac{2\sigma^2}{n}\left(I(W; Z^n) + \frac{M_m(W; Z^n)}{(\delta/2)^{1/m}} + \log\frac{2}{\delta}\right)}, \tag{32}$$

where $M_m(W; Z^n)$, defined in (9), is the mth root of the mth central moment of the information density.

Proof: We shall use Markov's inequality in the following form: for a random variable U,

$$P_U\left[U \leq \mathbb{E}[U] + \frac{\mathbb{E}^{1/m}[|U-\mathbb{E}[U]|^m]}{\delta^{1/m}}\right] \geq 1-\delta. \tag{33}$$

Applying (33) to the information density random variable, we conclude that, with probability at least $1-\delta$,

$$\imath(W, Z^n) \leq I(W; Z^n) + \frac{M_m(W; Z^n)}{\delta^{1/m}}. \tag{34}$$

It now follows from (30), (34), and the union bound that, with probability at least $1-2\delta$,

$$|\text{gen}(W, Z^n)| \leq \sqrt{\frac{2\sigma^2}{n}\left(I(W; Z^n) + \frac{M_m(W; Z^n)}{\delta^{1/m}} + \log\frac{1}{\delta}\right)}. \tag{35}$$

The desired result follows after the substitution $\delta \to \delta/2$. ■

Corollary 7: Under the assumptions in Theorem 1, the following bound holds with probability at least $1-\delta$ under P_{WZ^n}:

$$|\text{gen}(W, Z^n)| \leq \sqrt{\frac{2\sigma^2}{n}\left(\mathcal{L}(Z^n \to W) + 2\log\frac{2}{\delta}\right)}. \tag{36}$$

Here, $\mathcal{L}(Z^n \to W)$ denotes the maximal leakage, defined as

$$\mathcal{L}(Z^n \to W) = \log \mathbb{E}_{P_W}\left[\operatorname*{ess\,sup}_{P_{Z^n}} \frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}\right]. \tag{37}$$

Proof: Markov's inequality implies that, with probability at least $1-\delta$ under P_{WZ^n},

$$\imath(W, Z^n) \leq \log \mathbb{E}_{P_{WZ^n}}\left[\frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}\right] + \log\left(\frac{1}{\delta}\right) \tag{38}$$

Next, we can bound the expectation over $P_{Z^n|W}$ by an essential supremum:

$$\mathbb{E}_{P_W P_{Z^n|W}}\left[\frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}\right] \leq \mathbb{E}_{P_W}\left[\operatorname*{ess\,sup}_{P_{Z^n|W}} \frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}\right]. \tag{39}$$

The assumption that $P_{WZ^n} \ll P_W P_{Z^n}$ means that any set in the support of P_{WZ^n} is also in the support of $P_W P_{Z^n}$. We can therefore upper-bound the ess sup as follows:

$$\operatorname*{ess\,sup}_{P_{Z^n|W}} \frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}} \leq \operatorname*{ess\,sup}_{P_{Z^n}} \frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}. \tag{40}$$

Combining (38)-(40), we see that

$$\imath(W, Z^n) \leq \log \mathcal{L}(Z^n \to W) + \log\left(\frac{1}{\delta}\right), \tag{41}$$

which, combined with (30) through the union bound and the substitution $\delta \to \delta/2$, gives the desired result. ■

D. Remarks on the Tail Bounds in Sections II-B and II-C

The single-draw tail bound in (32) reveals a relation between the central moments of the information density and the confidence parameter δ. Specifically, the higher the moment of the information density that can be controlled, the more benign the dependence of the generalization error bound on δ. A similar observation holds for the data-independent PAC-Bayesian bound (26), in which controlling higher moments of the random variable $D(P_{W|Z^n} \,||\, P_W)$ leads to a more favorable dependence of the generalization bound on δ.

In the limit $m \to \infty$ the bound in (32) reduces to

$$|\text{gen}(W, Z^n)| \leq \sqrt{\frac{2\sigma^2}{n}\left(I(W; Z^n) + M_\infty(W; Z^n) + \log\frac{2}{\delta}\right)}, \tag{42}$$

where $M_\infty(W; Z^n) = \operatorname{ess\,sup}_{P_{WZ^n}} |\imath(w, z^n) - I(W; Z^n)|$. So, in this limit, the dependence on δ is of order $\sqrt{\log(1/\delta)}$. However, the bound (36) is tighter than (42), up to the factor 2 multiplying the logarithm. It is also tighter than the max information bound in [14, Thm. 4] with $\beta = 0$, up to the aforementioned factor of 2. Indeed, let the max information be defined as

$$I_{\max}(W; Z^n) = \operatorname*{ess\,sup}_{P_{WZ^n}} \imath(w, z^n). \tag{43}$$

[2] Note that the argument of the square root can be negative, but that this happens with probability at most δ. Therefore, the right-hand side of (30) is well-defined with probability at least $1-\delta$.

It is readily verified that

$$I_{\max}(W;Z^n) \leq I(W;Z^n) + M_\infty(W;Z^n). \tag{44}$$

As shown in [5, Lem. 12], $\mathcal{L}(Z^n \to W) \leq I_{\max}(W;Z^n)$. Thus, provided that

$$\mathcal{L}(Z^n \to W) \leq I_{\max}(W;Z^n) + \log\frac{2}{\delta}, \tag{45}$$

we have established that the bound in (36) is stronger than, in order, the max information bound in [14, Thm. 4] with $\beta = 0$, and (42). However, the maximal leakage bound in [5, Cor. 10] is still stronger than the one in (36) by a $\log 2/\delta$ term inside the square root.

In the next section, we present a different approach to obtaining single-draw tail bounds, which reveals a coupling between δ and the tail of the information density random variable.

III. Bounds via the Strong Converse

As pointed out in Section I, a key tool for deriving the single-draw bound (5) is the data processing inequality for f-divergences. This is also true for some of the bounds presented in [5]. In the context of binary hypothesis testing, it is known that such an inequality only leads to a weak converse bound on the region of achievable error rates. To obtain a strong converse, one needs to use [13, Lem. 12.2] (restated in Lemma 8 below for convenience), which provides a bound on the probability of an event under a distribution P in terms of its probability under Q.

Lemma 8: Let E be an arbitrary event and P and Q be probability measures such that P is absolutely continuous with respect to Q. Then, for all $\gamma \in \mathbb{R}$,

$$P[E] \leq P\left[\log\frac{\mathrm{d}P}{\mathrm{d}Q} > \gamma\right] + e^\gamma Q[E]. \tag{46}$$

As we shall show next, this inequality can be turned into a generalization bound by choosing P, Q, and E appropriately.

Theorem 9: Under the assumptions of Theorem 1, the following bound holds with probability at least $1-\delta$ over P_{WZ^n}:

$$|\mathrm{gen}(W,Z^n)| \leq \sqrt{\frac{2\sigma^2}{n}\left(\gamma + \log\left(\frac{2}{\delta - P_{WZ^n}[\imath(W,Z^n)\geq\gamma]}\right)\right)} \tag{47}$$

for all γ for which the arguments of the logarithm and the square root are nonnegative.

Proof: With $P = P_{WZ^n}$, $Q = P_W P_{Z^n}$ and

$$E = \{(w,z^n) : |\mathrm{gen}(w,z^n)| > \epsilon\}, \tag{48}$$

we apply Lemma 8 to get

$$P_{WZ^n}[E] \leq P_{WZ^n}\left[\imath(W,Z^n) \geq \gamma\right] + e^\gamma P_W P_{Z^n}[E]. \tag{49}$$

The σ-subgaussianity of the loss function implies that [10, Eq. (2.9)]

$$P_{Z^n}[|\mathrm{gen}(w,Z^n)| > \epsilon] \leq 2\exp\!\left(-n\epsilon^2/(2\sigma^2)\right). \tag{50}$$

Inserting (50) into (49), we obtain

$$\begin{aligned} &P_{WZ^n}[|\mathrm{gen}(W,Z^n)| > \epsilon] \\ &\quad \leq P_{WZ^n}[\imath(W,Z^n)\geq\gamma] + 2\exp\!\left(\gamma - n\epsilon^2/(2\sigma^2)\right). \end{aligned} \tag{51}$$

We get the desired result by imposing that the right-hand side of (51) is less than δ and solving for ϵ. ■

Unlike the bounds in Section II, this bound depends on the tail distribution of the information density. For a given δ, the parameter γ needs to be chosen large enough to make the factor $\delta - P_{WZ^n}[\imath(W,Z^n)\geq\gamma]$ positive. However, choosing γ too large makes the bound loose because of the γ term that is added to the log. This reveals a trade-off between the rate of decay of the tail of the information density and the confidence level δ.

Controlling the tail of the information density results in a tighter bound than the moment-based bound in (32) and the maximal leakage bound [5, Cor. 10] (up to some $\log 1/\delta$ terms). Indeed, these two bounds can be obtained by further upper-bounding the right-hand side of (47), as we shall discuss next.

A. Moment-Based Single-Draw Tail Bound

By Markov's inequality,

$$\begin{aligned} &P_{WZ^n}[\imath(W,Z^n)\geq\gamma] \\ &\leq P_{WZ^n}[|\imath(W,Z^n) - I(W;Z^n)| \geq \gamma - I(W;Z^n)] \qquad (52)\\ &\leq \frac{(M_m(W;Z^n))^m}{(\gamma - I(W;Z^n))^m}. \qquad (53) \end{aligned}$$

We now set

$$\gamma = \frac{M_m(W;Z^n)}{(\delta/2)^{1/m}} + I(W;Z^n). \tag{54}$$

Subsituting (54) in (53), we conclude that

$$P_{WZ^n}[\imath(W,Z^n)\geq\gamma] \leq \delta/2. \tag{55}$$

Inserting this upper bound into (47) we obtain

$$\epsilon \leq \sqrt{\frac{2\sigma^2}{n}\left(I(W;Z^n) + \frac{M_m(W;Z^n)}{(\delta/2)^{1/m}} + \log\frac{4}{\delta}\right)}, \tag{56}$$

which coincides with (32), up to a $\log 2$ term.

B. Maximal Leakage Single-Draw Tail Bound

Using the assumption that $P_{WZ^n} \ll P_W P_{Z^n}$, we get

$$P_{WZ^n}[\imath(W,Z^n)\geq\gamma] \leq P_W\left[\operatorname*{ess\,sup}_{P_{Z^n}} \frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}} \geq e^\gamma\right]. \tag{57}$$

Thus, Markov's inequality implies that

$$P_{WZ^n}[\imath(W,Z^n)\geq\gamma] \leq e^{-\gamma}\,\mathbb{E}_{P_W}\left[\operatorname*{ess\,sup}_{P_{Z^n}} \frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}\right]. \tag{58}$$

Setting $\gamma = \mathcal{L}(Z^n \to W) + \log(2/\delta)$ and using this result in (47), we get, with probability at least $1-\delta$ over P_{WZ^n},

$$|\mathrm{gen}(W,Z^n)| \leq \sqrt{\frac{2\sigma^2}{n}\left(\mathcal{L}(Z^n \to W) + \log\frac{4}{\delta} + \log\frac{2}{\delta}\right)}. \tag{59}$$

References

[1] D. Russo and J. Zou, “Controlling Bias in Adaptive Data Analysis Using Information Theory,” in *Artificial Intelligence and Statistics*, May 2016, pp. 1232–1240.

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

[3] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, “Learners that Use Little Information,” *Proc. Algorithmic Learning Theory, PLMR*, vol. 83, no. 25-55, 2018.

[4] Y. Bu, S. Zou, and V. V. Veeravalli, “Tightening Mutual Information Based Bounds on Generalization Error,” Jan. 2019, arXiv: 1901.04609.

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

[6] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: from theory to algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[7] D. A. McAllester, “Some PAC-bayesian theorems,” in *Proc. Conf. Computational Learning Theory (COLT)*, Jul. 1998, pp. 230–234.

[8] B. Guedj, “A Primer on PAC-Bayesian Learning,” Jan. 2019, arXiv. [Online]. Available: http://arxiv.org/abs/1901.05353

[9] O. Catoni, “PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning,” *IMS Lecture Notes Monogr. Ser.*, vol. 56, pp. 1–163, 2007.

[10] M. J. Wainwright, *High-dimensional statistics: a nonasymptotic viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

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

[12] I. Issa, S. Kamath, and A. B. Wagner, “An operational measure of information leakage,” in *2016 Annual Conference on Information Science and Systems (CISS)*, March 2016, pp. 234–239.

[13] Y. Polyanskiy and Y. Wu, *Lecture Notes On Information Theory*, Cambridge, U.K., 2019.

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