PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes

Peter Grünwald[*] Thomas Steinke[†] Lydia Zakynthinou[‡]

June 18, 2021

Abstract

We give a novel, unified derivation of *conditional* PAC-Bayesian and mutual information (MI) generalization bounds. We derive conditional MI bounds as an instance, with special choice of prior, of conditional *MAC*-Bayesian (Mean Approximately Correct) bounds, itself derived from conditional PAC-Bayesian bounds, where 'conditional' means that one can use priors conditioned on a joint training and ghost sample. This allows us to get nontrivial PAC-Bayes and MI-style bounds for general VC classes, something recently shown to be impossible with standard PAC-Bayesian/MI bounds. Second, it allows us to get faster rates of order $O((\mathsf{KL}/n)^{\gamma})$ for $\gamma > 1/2$ if a Bernstein condition holds and for exp-concave losses (with $\gamma = 1$), which is impossible with both standard PAC-Bayes generalization and MI bounds. Our work extends the recent work by Steinke and Zakynthinou [2020] who handle MI with VC but neither PAC-Bayes nor fast rates, the recent work of Hellström and Durisi [2020] who extend the latter to the PAC-Bayes setting via a unifying exponential inequality, and Mhammedi et al. [2019] who initiated fast rate PAC-Bayes generalization error bounds but handle neither MI nor general VC classes.

1 Extended Introduction

We first give a mini-introduction to PAC-Bayesian and mutual information bounds. Then we indicate two deficiencies of such bounds and give an informal statement of our main result, which solves both issues for both types of bounds at the same time. At the end of the introduction we discuss related work. In the remaining sections 2–4, we provide additional mathematical preliminaries, then we state our main lemma (proof delegated to an appendix) and use it to prove our main theorem and present its implications.

Setting In the standard setting of supervised learning, we are given a *model*, i.e., a set $\mathcal{F}$, where each $f \in \mathcal{F}$ is a hypothesis that takes the form of a *predictor*. Our aim is to learn to predict well based on a sample of n i.i.d. examples $Z = (Z_1, \ldots, Z_n)$ drawn from an unknown distribution $\mathcal{D}$ over the space of examples, $\mathcal{Z}$. We will denote the random variable representing a sample by Z, whereas a single example will be denoted by a Z_i, as previously, or by Z'. We adopt the convention of using upper-case letters for random variables (RVs) and lower-case letters for their realizations. A *learning algorithm* $A : \mathcal{Z}^n \to \Delta(\mathcal{F})$ (where $\Delta(\mathcal{F})$ is the set of distributions over $\mathcal{F}$) takes as input the sample Z and outputs a distribution over hypotheses. The special case of deterministic predictors such as ERM is covered by allowing the algorithm to output distributions on a single $f \in \mathcal{F}$. We refer to the posterior distribution of the output of A given input Z by $A|Z$. For a loss function $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$, $\ell(f; z')$ denotes the loss of a deterministic hypothesis $f \in \mathcal{F}$ on an example $z' \in \mathcal{Z}$. We extend this notation to define the true loss and the empirical loss of f on a sample $z \in \mathcal{Z}^n$ by $\ell(f; \mathcal{D}) = \mathbb{E}_{Z' \sim \mathcal{D}}[\ell(f; Z')]$ and $\ell(f; z) = \frac{1}{n} \sum_{i=1}^{n} \ell(f; z_i)$, respectively. Furthermore, for a randomized hypothesis $F \in \Delta(\mathcal{F})$, we define the expected true loss and the empirical loss on sample $z \in \mathcal{Z}^n$ by $L(F; \mathcal{D}) = \mathbb{E}_{f \sim F}[\ell(f; \mathcal{D})]$ and $L(F; z) = \mathbb{E}_{f \sim F}[\ell(f; z)]$, respectively. A *learning problem* is a tuple $(\mathcal{D}, \ell, \mathcal{F})$.

[*] CWI Amsterdam and Leiden University. peter.grunwald@cwi.nl

[†] Google Research, Brain Team. fast@thomas-steinke.net

[‡] Khoury College of Computer Sciences, Northeastern University. zakynthinou.l@northeastern.edu

Accepted for publication at the 34th Annual Conference on Learning Theory (COLT 2021).

Standard PAC-Bayesian bounds Within this setting, a standard goal is to bound the *generalization error* of an algorithm A in terms of its *empirical/training error.* A standard way to achieve this, which has recently received renewed attention, are *PAC-Bayesian generalization error bounds* [McAllester, 1998, 2003, Langford and Shawe-Taylor, 2003, Seeger, 2002, Maurer, 2004, Audibert, 2004, Catoni, 2007, Ambroladze et al., 2007] which commonly take the form:

$$\overbrace{L(A|Z;\mathcal{D})}^{\text{generalization error}} - \overbrace{L(A|Z;Z)}^{\text{training error}} \trianglelefteq C_1 \cdot \sqrt{\frac{L(A|Z;Z)\cdot \mathsf{KL}(A|Z\|\pi)}{n}} + C_2 \cdot \frac{\mathsf{KL}(A|Z\|\pi)}{n} \tag{1}$$

for some constants $C_1, C_2 > 0$ and $\mathsf{KL}(A|Z\|\pi)$ being the KL divergence between the 'posterior' output of the algorithm and the 'prior' distribution π over $\mathcal{F}$. The bounds hold for arbitrary priors π, as long as these are chosen independently of the data Z. Here we are ignoring $O(\log n)$ factors. The notation $\trianglelefteq$ expresses that the equation holds up to a small additive term with high probability over the distribution $\mathcal{D}^n$ of the training sample Z as well as in expectation. To be precise, (1) holds as an *exponential stochastic inequality* or ESI (pronounced 'easy'), a useful concept introduced and used by Koolen et al. [2016] and Grünwald and Mehta [2020], which we will use throughout this paper.

Definition 1 (Exponential Stochastic Inequality (ESI) [Grünwald and Mehta, 2020])**.** *Let $\eta > 0$ and X, Y be random variables that can be expressed as functions of the random variable U defined on the probability space $\mathcal{D}^n$. Then*

$$X \trianglelefteq_{\eta}^{U} Y \Leftrightarrow \mathop{\mathbb{E}}_{U}\left[e^{\eta(X-Y)}\right] \leq 1.$$

When no ambiguity can arise, we omit the random variable U. Besides simplifying notation, ESIs are useful in that they simultaneously capture "with high probability" and "in expectation" results, that is, $X \trianglelefteq_{\eta}^{U} Y$, implies both that $\forall \delta \in (0,1)$, $X \leq Y + \log(1/\delta)/\eta$, with probability at least $1-\delta$ over the randomness of U and that $\mathbb{E}_U[X] \leq \mathbb{E}_U[Y]$.

The standard PAC-Bayes bound (1) has recently been applied to practically important continuously parameterized model classes, such as deep neural networks [Dziugaite and Roy, 2017, Zhou et al., 2019]. The prior then takes the form of a probability density over the parameters (e.g. weights $\vec{w}$) and in order for the KL term to be finite, one needs to randomize the output of the algorithm. Therefore, even if the empirical error of the output $\vec{w}|Z$ of the original learning algorithm (typically SGD) can be driven down to 0, the empirical error as appearing in (1), and therefore also the multiplication factor inside the square root, is not 0—one typically takes a Gaussian around $\vec{w}|Z$ leading to a nonnegligible $L(A|Z;Z)$ (Mhammedi et al. [2019] provide a numerical example).

Standard Mutual Information (MI) Bounds Another, related way to bound generalization error is provided by *mutual information bounds* [Russo and Zou, 2016, Xu and Raginsky, 2017]. These usually take on the following form:

$$\left|\mathop{\mathbb{E}}_{Z}[L(A|Z;\mathcal{D}) - L(A|Z;Z)]\right| \leq \sqrt{\frac{2\cdot I(A|Z;Z)}{n}}, \tag{2}$$

with $I(A|Z;Z)$ denoting the mutual information between the training data and the algorithm's output.

Two Issues with the Bounds Standard PAC-Bayesian and MI bounds have two deficiencies in common. First, as recently shown by Livni and Moran [2020], there exist hypothesis classes with finite Vapnik-Chervonenkis (VC) dimension d for which, rather than achieving the standard VC generalization error bound of order $\sqrt{(d\log n)/n}$, PAC-Bayes bounds of the form (1) must remain trivial: there exists a VC class, such that for any arbitrary learning algorithm A, there exists a realizable (i.e., $\inf_{f\in\mathcal{F}} \ell(f;\mathcal{D}) = 0$) distribution $\mathcal{D}$, such that for any prior π (even one that is allowed to depend on the data-generating distribution $\mathcal{D}$), either the KL divergence term $\mathsf{KL}(A|Z\|\pi)$ is arbitrarily large or the loss is large ($L(A|Z;\mathcal{D}) > 1/4$). Similarly, Bassily et al. [2018] and Nachum et al. [2018] show that there exists a VC class such that, for any proper and consistent learning algorithm A, there exists a realizable distribution $\mathcal{D}$, such that the mutual information $I(A|Z;Z)$ in the bound of (2) is arbitrarily large.

Second, in both theoretically interesting settings (such as random label noise, see Example 1 below) and in practical settings (as already indicated above) the empirical error term $L(A|Z;Z)$ inside the square root of (1) often cannot be ignored. Then both bounds (1) and (2) will be of order $\sqrt{\textsc{complexity}/n}$. The theory of *excess risk bounds* suggests that this is, in many cases, suboptimal and we can obtain a more desirable bound of the form $\textsc{complexity}/n$. Here we concentrate on the following typical form of PAC-Bayesian excess risk bounds [Audibert, 2004, Zhang, 2006a,b, Grünwald and Mehta, 2020, 2019], but the results are comparable in nature to excess risk bounds based on e.g. Rademacher complexity bounds [Bartlett and Mendelson, 2006]:

$$\overbrace{R(A|Z;\mathcal{D})}^{\text{excess risk}} \trianglelefteq C_3 \cdot \overbrace{R(A|Z;Z)}^{\text{empirical excess risk}} + C_4 \cdot \left(\frac{\mathsf{KL}(A|Z\|\pi)}{n}\right)^{\gamma} \tag{3}$$

for some constants $C_3, C_4 > 1$ and $\gamma \in [1/2, 1]$. Here we ignore $O(\log\log n)$ factors. The *excess risk* of a distribution over predictors $F \in \Delta(\mathcal{F})$ is defined as $R(F;\mathcal{D}) = L(F;\mathcal{D}) - L(f^*;\mathcal{D})$ where f^* is an optimal predictor within the class $\mathcal{F}$, achieving $\min_{f\in\mathcal{F}} \ell(f;\mathcal{D})$, whose existence is commonly assumed (e.g. Tsybakov [2004], Bartlett and Mendelson [2006], Grünwald and Mehta [2020]). The excess risk of algorithm A based on training sample Z, $R(A|Z;\mathcal{D})$, is thus a nonnegative random variable (depending on Z) denoting the additional risk incurred if one predicts based on the learned distribution $A|Z$, compared to the best one could have with knowledge of the true distribution $\mathcal{D}$. Similarly, the *empirical excess risk* of F on a sample $z \in \mathcal{Z}^n$ is $R(F;z) = L(F;z) - L(f^*;z)$. Substituting these terms and rearranging, inequality (3) can be written as follows, giving an upper bound on the generalization gap:

$$L(A|Z;\mathcal{D}) - L(A|Z;Z) \trianglelefteq (L(f^*;\mathcal{D}) - L(f^*;Z)) + (C_3 - 1)\cdot R(A|Z;Z) + C_4 \cdot \left(\frac{\mathsf{KL}(A|Z\|\pi)}{n}\right)^{\gamma} \tag{4}$$

The γ for which (3) holds depends on the interplay between the model $\mathcal{F}$, the loss function ℓ, and the true distribution $\mathcal{D}$. Specifically, a sufficient condition for the result to hold for $\gamma = 1/(2-\beta)$ is if the learning problem $(\mathcal{D}, \ell, \mathcal{F})$ satisfies a (B,β)-*Bernstein condition* [Bartlett et al., 2002, Bartlett and Mendelson, 2006, Van Erven et al., 2015]:

Definition 2 (Bernstein Condition). *Let $\beta \in [0,1]$ and $B \geq 1$. Then $(\mathcal{D}, \ell, \mathcal{F})$ satisfies the (B,β)-*Bernstein condition *if there exists a $f^* \in \mathcal{F}$ such that*

$$\mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}\left[(\ell(f;Z') - \ell(f^*;Z'))^2\right] \leq B\Big(\mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}[\ell(f;Z') - \ell(f^*;Z')]\Big)^{\beta} \quad \textit{for all } f \in \mathcal{F}. \tag{5}$$

If the Bernstein condition (5) holds for some f^*, then this f^* must be an optimal predictor as above. If the losses are assumed bounded then the Bernstein condition vacuously holds for $\beta = 0$ with some B. Throughout this paper, the losses are assumed in $[0,1]$, hence it always holds with $\beta = 0, B = 4$. Therefore, the *slow rate* of $\gamma = 1/(2-0) = 1/2$ can always be obtained. But for loss functions with curvature (specifically, all bounded so-called *mixable* loss functions, which includes all *exp-concave* loss functions [Van Erven et al., 2015]), the Bernstein condition also holds with $\beta = 1$, implying *fast* $O(1/n)$ rates, i.e., $\gamma = 1$. Examples include the bounded squared error loss and logistic loss. Specifically, for the squared loss $\ell(f;(X,Y)) := (Y - f(X))^2$ (rescaled so that all functions map X to $[-1/2, 1/2]$ and $Y \in [-1/2, 1/2]$ so that the range is $[0,1]$) it automatically holds with $\beta = 1$ and $B = 4$ [Grünwald and Mehta, 2020, Proposition 19]. Even for the nonmixable 0/1-loss, a Bernstein condition may still hold. For example, in the realizable case and in the case of random label noise (homoskedasticity), the Massart condition and, hence, the Bernstein condition holds, giving $\gamma = 1$. The Bernstein condition is a significant weakening of the perhaps more well-known Tsybakov-Mammen [Tsybakov, 2004] condition which itself is a weakening of the Massart condition for classification; see Van Erven et al. [2015] for an extensive overview and links between a variety of "easiness" conditions such as (Massart, Bernstein and Tsybakov) proposed in the literature. Tsybakov [2004] provides examples of situations in which Bernstein holds for β strictly between 0 and 1, where *faster/intermediate* rates can be obtained.

For many algorithms, the empirical excess risk term $R(A|Z;Z)$ will be negligible. For example, for ERM (Empirical Risk Minimization) it will automatically be nonpositive since by definition the ERM cannot have larger loss on the sample than f^*. In addition, the first term, that is, the excess risk of f^*, disappears when the inequality is weakened to an in-expectation bound, while introducing a small unavoidable term in the in-probability bound. Then, in many settings, the right-hand side of (4) is clearly smaller than that of (1) which suggests that the standard generalization bound (1) is suboptimal as soon as a Bernstein condition holds with $\beta > 0$. Below we shall see that this is indeed the case.

Solving Both Issues at Once for both Bounds Partial solutions for both issues were provided by Audibert [2004], Catoni [2007], Mhammedi et al. [2019], Steinke and Zakynthinou [2020], Hellström and Durisi [2020]. By combining their insights and adding a new fundamental lemma (Lemma 1 below), we manage to solve both problems for both types of bounds in essentially a single derivation. Its first intermediate conclusion is the following *faster rate data-conditional generalization error bound* (Theorem 1 below): Let $(\mathcal{D}, \ell, \mathcal{F})$ represent a learning problem which satisfies the (B, β)-Bernstein condition and suppose the loss function ℓ is bounded. Let the data $\tilde{Z}_{\mathbf{0}} = (\tilde{Z}_{1,0}, \dots, \tilde{Z}_{n,0})^\top \in \mathcal{Z}^n$ be i.i.d. $\sim \mathcal{D}$. Then for arbitrary *almost exchangeable data-dependent priors* $\pi \mid \langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}} \rangle$ we have:

$$L(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) - L(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}}) \;\trianglelefteq\; (1 \wedge 2\beta)\cdot R(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}}) + O\left(\frac{\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}\rangle\right)\right]}{n}\right)^{\frac{1}{2-\beta}} + \frac{6\eta}{n} \tag{6}$$

Here $\wedge$ denotes minimum, the result holds up to $\log\log n$ factors and it requires an additional condition which essentially holds as long as $\mathsf{KL}(\cdot\|\cdot) = o(n)$ almost surely under $\mathcal{D}$. Note that this is an ESI inequality and as such it holds both in expectation and up to a small additive term with high probability over the training sample $\tilde{Z}_{\mathbf{0}}$. We return later to this fact and to the remainder term $6\eta/n$, which for now may be thought of as negligible.

To appreciate (6), first note that, since the Bernstein condition automatically holds for $\beta = 0$, so does (6). Then the first term on the right disappears and the KL term becomes of order $\sqrt{\mathsf{KL}/n}$, as is the leading term for classical PAC-Bayesian bounds. However, in stark contrast to classical PAC-Bayesian bounds, we are now allowed (not required) to use priors which can *depend on the data in many – but not arbitrary – ways*: just like in classical Vapnik-Chervonenkis learning theory, we imagine a *ghost sample* $\tilde{Z}_{\mathbf{1}}$ of equal size and distribution as the training sample $\tilde{Z}_{\mathbf{0}}$. The notation

$$\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}} \rangle := (\{\tilde{Z}_{1,0}, \tilde{Z}_{1,1}\}, \{\tilde{Z}_{2,0}, \tilde{Z}_{2,1}\}, \dots, \{\tilde{Z}_{n,0}, \tilde{Z}_{n,1}\})^\top$$

indicates a vector of n *unordered* pairs of examples, where the i-th component is the bag of example i in the training sample $\tilde{Z}_{\mathbf{0}}$ and example i in the ghost sample $\tilde{Z}_{\mathbf{1}}$. The prior $\pi|\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}} \rangle$ is allowed to depend on these $2n$ examples that include all the n training examples, but all information as to whether an example is in the training or ghost sample is hidden from the prior. The complexity is then measured as the *expected* KL divergence where the ghost sample is i.i.d. $\sim \mathcal{D}$. More formally, let us write $\tilde{Z}_S = (\tilde{Z}_{1,S_1}, \dots, \tilde{Z}_{n,S_n})^\top \in \mathcal{Z}^{n \times 1}$ for the sample whose i-th example belongs to the sample $\tilde{Z}_{\mathbf{0}}$ or $\tilde{Z}_{\mathbf{1}}$, as indicated by $S_i \in \{0, 1\}$ and let $\tilde{Z}_{\bar{S}} = (\tilde{Z}_{1,\bar{S}_1}, \dots, \tilde{Z}_{n,\bar{S}_n})^\top$ be its complement.

Definition 3 (Almost Exchangeable Prior, terminology from Audibert [2004])**.** *A function (conditional distribution)* $\pi : \mathcal{Z}^{n\times 2} \to \Delta(\mathcal{F})$ *is* almost exchangeable *if for all* $\tilde{z} \in \mathcal{Z}^{n\times 2}$*, it holds that* $\pi|(\tilde{z}_s, \tilde{z}_{\bar{s}}) = \pi|(\tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}})$*,* $\forall s \in \{0,1\}^n$*, justifying the notation* $\pi|\langle \tilde{z}_s, \tilde{z}_{\bar{s}} \rangle = \pi|\langle \tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}} \rangle$*.*

It may appear that the expectation over the ghost sample makes such KL bounds incalculable in practice, but this is not so: in Section 3.1 we give examples of data-dependent almost exchangeable priors for which the KL complexity term, or at least a good upper bound, can be calculated based on the observed data. In particular, in classification with a class $\mathcal{F}$ with finite VC dimension d, when an ERM algorithm with a specific consistency property is used (Theorem 2 shows that such an ERM can always be constructed), the KL term can be bounded as $d\log(2n)$, leading us to recover classical VC bounds; similarly, for size k-compression schemes, the KL term is also bounded as $k\log(2n)$.

Now suppose a Bernstein condition holds for some $\beta > 0$. We then see that (6) gives a *faster rate bound* of the same flavour as the classical PAC-Bayesian excess risk bound (4), and with the same exponent γ. In particular, if ERM is used then the excess risk term will be nonpositive and only the faster-rate term remains. We also provide a class of exchangeable priors for which a Gibbs posterior can be calculated based on the observed data, and for the corresponding Gibbs predictor we also get a bound in which the excess risk term can be omitted (Example 2).

Note that the empirical excess risk term in excess risk bounds does not necessarily vanish if $\beta \downarrow 0$: the RHS of our result (6) provides the best of the RHS of (4) and (1). For ERM, if the best β in the Bernstein condition is known (e.g., for bounded squared or logistic loss), the bound (6) is empirical—it can be calculated from the data only. If, as in classification, we do not know the best β in advance, or we do not use ERM so that the R term is hard to quantify without knowing f^*, the bound as such cannot be calculated based on the data only; we return to this issue in Section 4.

We may view both the algorithm A and the data-dependent prior π as *conditional* distributions over $\mathcal{F}$, given the training sample $\tilde{Z}_{\mathbf{0}}$, and the vector of unordered pairs $\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}} \rangle$, respectively. Of course, when designing the prior π we can also take into account the algorithm A: given $\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}} \rangle$, there are only 2^n outputs possible for any deterministic algorithm A (such as ERM) that outputs a single distribution given training sample $\tilde{Z}_{\mathbf{0}}$. As an additional benefit, we can thus take, without loss of generality, a prior π with discrete support of at most 2^n elements, allowing us to provide bounds for general nonrandomized learning algorithms – something which, as we have already seen, is not possible in the standard PAC-Bayesian setup when the parameters of $\mathcal{F}$ are continuous-valued.

Solving both Issues for Mutual Information As mentioned above, our bound (6) holds as an *exponential stochastic inequality* (Definition 1). Formally, an ESI has the following implications.

Proposition 1 (ESI Implications [Mhammedi et al., 2019, Prop.9])**.** *If $X \trianglelefteq_\eta Y$, then $\forall \delta \in (0,1)$, $X \leq Y + \frac{\log \frac{1}{\delta}}{\eta}$, with probability at least $1-\delta$. Now let $\bar{\eta} > 0$ and let $g : [0, \bar{\eta}]$ be continuous and nondecreasing. If for all η with $0 < \eta \leq \bar{\eta}$, $X \trianglelefteq_\eta Y + g(\eta)$, then $\mathbb{E}[X] \leq \mathbb{E}[Y] + g(0)$.*

Our main Theorem 1, rendered as (6) above, holds with $\trianglelefteq$ instantiated to $\trianglelefteq_\eta$ with every $0 < \eta \leq c\sqrt{n}$ for some constant $c > 0$. It can thus be weakened, by applying the proposition above with $g(\eta) = 6\eta/n$, to an in-probability PAC statement (setting $\eta = c\sqrt{n}$, it holds with probability at least $1-\delta$ up to $6c/\sqrt{n} + (-\log \delta)/(c\sqrt{n}) = O(1/\sqrt{n})$) but also to an in-expectation statement in which the remainder term $6\eta/n$ disappears. We then get a *MAC*-Bayesian bound, with MAC standing for 'Mean Approximately Correct'. By plugging into (6) a special almost exchangeable prior that is both distribution- and data-dependent, namely the prior that minimizes the bound in expectation for the given learning algorithm, we get the corresponding *faster-rate conditional mutual information bound*:

$$\underset{\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[L(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) - L(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}})\right] \leq (1 \wedge 2\beta) \cdot \underset{\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[R(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}})\right] + O\left(\frac{\mathsf{CMI}_{\mathcal{D}}(A)}{n}\right)^{\frac{1}{2-\beta}} \tag{7}$$

The term $\mathsf{CMI}_{\mathcal{D}}(A) = \inf_\pi \underset{\tilde{Z}_{\mathbf{0}},\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}\rangle\right)\right]$ denotes the *conditional mutual information of A* with respect to data distribution $\mathcal{D}$, introduced by Steinke and Zakynthinou [2020] as an information complexity measure, which is always finite, avoiding the impossibility results of Bassily et al. [2018]. This conditioning approach has already proven useful in proving sharper generalization bounds [Haghifam et al., 2020]. However, until the present work, no *fast rate* results had been proven with respect to CMI.

In contrast to the standard bound (2), there are no absolute signs on the left, but this is not of great concern since we are almost always interested in a one-sided bound anyway. If $\beta = 0$, the right-hand side of the bound (7) is smaller than that of (2), since $\mathsf{CMI}_{\mathcal{D}}(A) \leq I(A|Z;Z)$ [Haghifam et al., 2020]. Under a Bernstein condition or bounded loss with curvature, where $\beta > 0$, the rate is clearly faster than the rate obtained by the standard CMI bound, albeit with an additional excess risk term. For ERM, this first term disappears, and more generally in most interesting settings, the complexity term is the dominant term.

In Expectation vs. In Probability – A Paradox? At first sight, a fast rate means 'with high probability, convergence happens at rate faster than $O(1/\sqrt{n})$'. But this is impossible even in trivial cases with $\mathcal{F} = \{f\}$ containing only one element (so every learning algorithm must output f, no matter what data are observed – there is no learning/overfitting): if $\ell(f, Z_1)$ has variance σ^2, then we find by the central limit theorem that for every fixed $\alpha < 1$, for all large n,

$$L(A|Z;\mathcal{D}) - L(A|Z;Z) = \underset{f\sim F}{\mathbb{E}}[\ell(f;\mathcal{D})] - \underset{f\sim F}{\mathbb{E}}[\ell(f;Z)] \geq C_\alpha \frac{\sigma}{\sqrt{n}}$$

with probability α over the training sample Z and a constant $C_\alpha > 0$. Yet, (6) still provides faster rates in a weaker sense. To see this, note first that, being an ESI, it implies convergence in expectation; and then the $1/\sqrt{n}$ term is really not there (and the Central Limit Theorem does not hurt us) – so we do get a faster rate in expectation. Second, the largest subscript η for which the ESI holds is of order $\sqrt{n}$ – implying that we do incur $O(1/\sqrt{n})$-fluctuations, and do not contradict the central limit theorem. Yet importantly, the square-root term has been *decoupled* from the KL complexity term, which (if $\beta = 1$) can converge to 0 as fast as $O(\mathsf{KL}/n)$. In contrast, all other PAC-Bayes bounds we know of, except those of Mhammedi et al. [2019], have the KL/n term *inside* the square root. If the KL term grows with n, as it usually does, this may make the convergence rate of such classical bounds substantially worse than $O(1/\sqrt{n})$. Thus, borrowing the terminology of Mhammedi et al. [2019], we really have *faster rates in probability up to an* irreducible, complexity-free $O(1/\sqrt{n})$ *term.*[1]

1.1 Related Work; Other Extensions of the Standard PAC-Bayesian Equation

Although they sometimes look different, most PAC-Bayes bounds can, potentially after slight relaxation, be brought in the form (1). Examples include the well-known bound with KL on the left due to Langford and Shawe-Taylor [2003], Seeger [2002], Maurer [2004] and the standard bound due to Catoni [2007]; see also Tolstikhin and Seldin [2013], who provide an overview and discussion of this type of bound. Based on an empirical Bernstein analysis, Tolstikhin and Seldin [2013] replaced the empirical error term inside the square root in (1) by a smaller second order term which, however, still is close to 0 only when the empirical error itself is close to 0. Based on a variation of the empirical Bernstein idea, a lemma which they called *un-expected* Bernstein, Mhammedi et al. [2019] replace the empirical error term inside the square root by a different second-order term which, they show, goes to 0 with high probability whenever a Bernstein condition holds. Thus, they are presumably one of the first to have a fast rate PAC-Bayesian *generalization* error bound (note again that fast PAC-Bayes *excess* risk bounds have been known for a long time). Their Theorem 7 provides a first version of the in-probability version of our (6), but with the $(1 \wedge 2\beta)$ replaced by 1 and the empirical excess risk $R(A|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}})$ replaced by (essentially) three times the standard risk (i.e., expected loss difference), making their first term larger than ours and not converge to 0 for algorithms for which the excess risk does not converge to 0; also their analysis is based on priors that do not allow conditioning on a ghost sample. However, in contrast to our bound, their bound has the pleasant property of being fully empirical, a point to which we return in Section 4. Simultaneously, Yang et al. [2019] also gave a fast rate PAC-Bayes generalization bound using a different technique, which includes a so-called 'flatness' term attempting to capture the flatness of the empirical risk surface on which the posterior Gibbs classifier concentrates. If this term is small with high probability, then the bound converges fast. In contrast, our bound converges fast when the strong Bernstein condition ($\gamma = 1$) holds and achieves faster rates otherwise. It is easy to show the 'flatness' term of [Yang et al., 2019] can be large even if a strong Bernstein condition holds; on the other hand, there may also be cases in which their bound is tighter than ours — the bounds are so different that they are hard to compare in general.

The Other Type of Data-Dependent Prior Mhammedi et al. [2019] do make use of data-dependent priors, an idea pioneered by Ambroladze et al. [2007], which is to set aside part of the training data and condition everything on it. In the simplest instance, one uses the learning algorithm's output on the first half of the data as a prior, then performs a standard PAC-Bayesian bound such as (1) on the second half. In this way one looses a factor of 2 in the bound but gets a much better informed prior, making the final bound often substantially better in practice (e.g. in [Dziugaite et al., 2021]). Mhammedi et al. [2019] extend this idea to using both half samples and mixing the results, analogously to cross-validation. Note though that this is a very different type of data-dependency than ours: the prior is given the full first half of the sample, rather than the full training sample plus a ghost sample with ordering information removed.

The Core of Our Contribution *MAC*-Bayesian bounds, although not under that name, are already to be found in Catoni's monograph [Catoni, 2007]. Catoni already mentions that the prior that minimizes a MAC-Bayesian bound is the prior that turns the KL term into the mutual information. Moreover, Catoni [2007], as

[1] For ESI-excess risk bounds, because of the substraction of $\ell(f^*; Z)$ in the bounds, the variance of the excess risk $L(A|Z; \mathcal{D})$ goes to 0 under a Bernstein condition and fast rates without the $O(1/\sqrt{n})$ term are possible — indeed, if $\beta > 0$ then (3) holds for an η that goes to 0 slower than $1/\sqrt{n}$ (Grünwald and Mehta [2020] provide various examples) and one gets in-probability excess risk bounds without the irreducible $O(1/\sqrt{n})$ term.

well as Audibert [2004] in his Ph.D. thesis, introduce the expected KL complexity based on almost exchangeable priors conditioned on a supersample, but these are not connected to conditional mutual information as in our paper. Even more closely related, Hellström and Durisi [2020] introduced an exponential inequality which yields conditional PAC-Bayesian and in-expectation bounds. However, none of the previous works connects fast rates to the conditional case with almost exchangeable prior. This is the crucial contribution of the present paper, hinging on our main, and novel, technical Lemma 1, which allows us to get fast rates. Below the lemma we explain how it goes beyond earlier developments.

2 Preliminaries

Additional Notation For convenience, we include a glossary with all frequently used symbols in Appendix A. For a random variable X and a distribution $\mathcal{P}$, we write $X \sim \mathcal{P}$ to denote that X is drawn from $\mathcal{P}$ and $X \sim \mathcal{P}^n$ to denote that X consists of n i.i.d. draws from $\mathcal{P}$. The distribution of a random variable X is denoted by $\mathcal{P}_X$ and will be omitted when it is clear from context. We denote the Bernoulli distribution over $\{0, 1\}$ with mean p by $\mathrm{Ber}(p)$. We also write $[n] = \{1, \dots, n\}$.

A supersample $\tilde{Z} = ((\tilde{Z}_{1,0}, \tilde{Z}_{1,1}), \dots, (\tilde{Z}_{n,0}, \tilde{Z}_{n,1}))^\top \sim \mathcal{D}^{n\times 2}$ is a vector of n pairs of i.i.d. examples, as in Table 1. Let $S = (S_1, \dots, S_n) \in \{0, 1\}^n$ such that $S \sim \mathrm{Ber}(1/2)^n$ and let $\bar{S}_i = 1 - S_i$ for all $i \in [n]$. We write $\tilde{Z}_S = (\tilde{Z}_{1,S_1}, \dots, \tilde{Z}_{n,S_n})^\top \in \mathcal{Z}^{n\times 1}$ for the sub-vector of $\tilde{Z}$ indexed by S and $\tilde{Z}_{\bar{S}} = (\tilde{Z}_{1,\bar{S}_1}, \dots, \tilde{Z}_{n,\bar{S}_n})^\top$ for its complement. Note that with this notation, we can write $\tilde{Z} = (\tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}})$, setting $S = \mathbf{0}$. We also refer to the vector of *unordered* pairs $\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}} \rangle = (\{\tilde{Z}_{1,0}, \tilde{Z}_{1,1}\}, \dots, \{\tilde{Z}_{n,0}, \tilde{Z}_{n,1}\})^\top$. With this notation, for any *almost exchangeable* prior distribution $\pi : \mathcal{Z}^{n\times 2} \to \Delta(\mathcal{F})$ (Definition 3) it holds that for all $\tilde{z} \in \mathcal{Z}^{n\times 2}$, $\forall s \in \{0, 1\}^n$, $\pi|\tilde{z} = \pi|(\tilde{z}_s, \tilde{z}_{\bar{s}}) = \pi|\langle \tilde{z}_s, \tilde{z}_{\bar{s}} \rangle = \pi|\langle \tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}} \rangle$.

$\tilde{Z}_{1,0}$	$\tilde{Z}_{1,1}$
$\tilde{Z}_{2,0}$	$\tilde{Z}_{2,1}$
$\vdots$	$\vdots$
$\tilde{Z}_{n,0}$	$\tilde{Z}_{n,1}$

Table 1: Supersample $\tilde{Z} \in \mathcal{Z}^{n\times 2}$

2.1 KL divergence and Mutual Information

First, we define the KL divergence of two distributions.

Definition 4 (KL Divergence). *Let $\mathcal{P}, \mathcal{Q}$ be two distributions over the space Ω and suppose $\mathcal{P}$ is absolutely continuous with respect to $\mathcal{Q}$. The* Kullback–Leibler (KL) divergence *(or* relative entropy*) from $\mathcal{Q}$ to $\mathcal{P}$ is*

$$\mathsf{KL}(\mathcal{P}\|\mathcal{Q}) = \mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[\log \frac{\mathcal{P}(X)}{\mathcal{Q}(X)}\right],$$

where $\mathcal{P}(X)$ and $\mathcal{Q}(X)$ denote the probability mass/density functions of $\mathcal{P}$ and $\mathcal{Q}$ on X, respectively.[2]

Next, we define mutual information.

Definition 5 (Mutual Information). *Let X, Y be two random variables jointly distributed according to $\mathcal{P}$. The mutual information of X and Y is*

$$I(X; Y) = \mathsf{KL}\big(\mathcal{P}_{(X,Y)}\big\|\mathcal{P}_X \times \mathcal{P}_Y\big) = \mathop{\mathbb{E}}_{X}\big[\mathsf{KL}\big(\mathcal{P}_{Y|X}\big\|\mathcal{P}_Y\big)\big],$$

where by $\mathcal{P}_X \times \mathcal{P}_Y$ we denote the product of the marginal distributions of $\mathcal{P}$ and $\mathcal{P}_{Y|X=x}(y) = \mathcal{P}_{(X,Y)}(x, y)/\mathcal{P}_X(x)$ is the conditional density function of Y given X.

[2] Formally, $\frac{\mathcal{P}(X)}{\mathcal{Q}(X)}$ is the Radon-Nikodym derivative of $\mathcal{P}$ with respect to $\mathcal{Q}$. If P is not absolutely continuous with respect to $\mathcal{Q}$ (i.e., $\frac{\mathcal{P}(X)}{\mathcal{Q}(X)}$ is undefined or infinite), then the KL divergence is defined to be infinite.

Definition 6 (Conditional Mutual Information)**.** *For random variables X, Y, Z, the mutual information of X and Y conditioned on Z is*

$$I(X; Y \mid Z) = I(X; (Y, Z)) - I(X; Z).$$

We define here the less common notion of *disintegrated mutual information*, as in [Negrea et al., 2019, Haghifam et al., 2020].

Definition 7 (Disintegrated Mutual Information)**.** *The* disintegrated mutual information *between X and Y given Z is*

$$I^Z(X; Y) = \mathsf{KL}\big(\mathcal{P}_{(X,Y)|Z} \big\| \mathcal{P}_{X|Z} \times \mathcal{P}_{Y|Z}\big),$$

where $\mathcal{P}_{(X,Y)|Z}$ denotes the conditional joint distribution of (X, Y) given Z and $\mathcal{P}_{X|Z}, \mathcal{P}_{Y|Z}$ denote the conditional marginal distributions of X, Y given Z, respectively.

The expected value of this quantity over Z is the Conditional Mutual Information between X and Y given Z that was defined above: $I(X; Y|Z) = \mathbb{E}_Z\big[I^Z(X; Y)\big]$.

We now define the Conditional Mutual Information of an Algorithm, as introduced in [Steinke and Zakynthinou, 2020].

Definition 8 (Conditional Mutual Information (CMI) of an Algorithm [Steinke and Zakynthinou, 2020])**.** *Let $A : \mathcal{Z}^n \to \Delta(\mathcal{F})$ be a randomized or deterministic algorithm. Let $\mathcal{D}$ be a probability distribution on $\mathcal{Z}$ and let $\tilde{Z} \in \mathcal{Z}^{n\times 2}$ be a supersample consisting of n pairs of examples, each example drawn independently from $\mathcal{D}$. Let $S \sim$ Ber$(1/2)^n$, independent from $\tilde{Z}$ and the randomness of A. Let $\tilde{Z}_S = (\tilde{Z}_{1,S_1}, \dots, \tilde{Z}_{n,S_n})^\top \in \mathcal{Z}^n$ – that is, $\tilde{Z}_S$ is the subset of $\tilde{Z}$ indexed by S.*

The conditional mutual information (CMI) of A with respect to $\mathcal{D}$ *is*

$$\mathsf{CMI}_{\mathcal{D}}(A) := I(A|\tilde{Z}_S; S \mid \tilde{Z}) = \mathop{\mathbb{E}}_{\tilde{Z}}\Big[I^{\tilde{Z}}(A|\tilde{Z}_S; S)\Big].$$

2.2 ESI Calculus

The following proposition is useful for our proofs.

Proposition 2 (ESI Transitivity and Chain Rule [Mhammedi et al., 2019, Prop.10])**.** *(a) Let $Z_1, \dots, Z_n$ be any random variables in $\mathcal{Z}$ (not necessarily independent). If for some $(\gamma_i)_{i\in[n]} \in (0, +\infty)^n$, $Z_i \trianglelefteq_{\gamma_i} 0$ for all $i \in [n]$, then*

$$\sum_{i=1}^n Z_i \trianglelefteq_{v_n} 0, \text{ where } v_n = \left(\sum_{i=1}^n \frac{1}{\gamma_i}\right)^{-1}.$$

(b) Suppose now that $Z_1, \dots, Z_n$ are independent and for some $\eta > 0$, for all $i \in [n]$, we have $Z_i \trianglelefteq_\eta 0$. Then $\sum_{i=1}^n Z_i \trianglelefteq_\eta 0$.

We now state a basic PAC-Bayesian result we use, under the ESI notation:

Proposition 3 (ESI PAC-Bayes [Mhammedi et al., 2019, Prop.11])**.** *Fix $\eta > 0$ and let $\{Y_f : f \in \mathcal{F}\}$ be any family of random variables such that for all $f \in \mathcal{F}$, $Y_f \trianglelefteq_\eta 0$. Let $\pi \in \Delta(\mathcal{F})$ be any distribution on $\mathcal{F}$ and let $A : \bigcup_{i=1}^n \mathcal{Z}^i \to \Delta(\mathcal{F})$ be a possibly randomized learning algorithm. Then*

$$\mathop{\mathbb{E}}_{f\sim A|Z}[Y_f] \trianglelefteq_\eta^Z \frac{\mathsf{KL}(A|Z\|\pi)}{\eta}.$$

Inside the proof of our main result we work with a random (*i.e.*, data-dependent) $\hat{\eta}$ in the ESI inequalities. We extend Definition 1 to this case by replacing the expectation in the definition of ESI by the expectation over the joint distribution of $(X,\ Y,\ \hat{\eta})$: $X \trianglelefteq_{\hat{\eta}} Y$ means that $\mathbb{E}[\exp(\hat{\eta}(X - Y))] \leq 0$. Via the following proposition one can tune η after seeing the data.

Proposition 4 (ESI from fixed to random η [Mhammedi et al., 2019, implied by Prop.12])**.** *Let $\mathcal{G}$ be a countable subset of $\mathbb{R}^+$ such that, for some $\eta_0 > 0$, for all $\eta \in \mathcal{G}$, $\eta \geq \eta_0$. Let π be a probability mass function over $\mathcal{G}$. Given a countable collection $\{Y_\eta : \eta \in \mathcal{G}\}$ of random variables satisfying $Y_\eta \trianglelefteq_\eta 0$, for all fixed $\eta \in \mathcal{G}$, we have, for arbitrary estimator $\hat{\eta}$ with support on $\mathcal{G}$,*

$$Y_{\hat{\eta}} \trianglelefteq_{\eta_0} \frac{-\log \pi(\hat{\eta})}{\hat{\eta}}.$$

2.3 Bernstein Condition

We consider learning problems $(\mathcal{D}, \ell, \mathcal{F})$ which satisfy the Bernstein Condition (Definition 2 in Section 1). It will be convenient to work with the following *linearized version* of the Bernstein condition, proven in Appendix B. It extends a well-known result that has appeared in previous work, e.g. in [Koolen et al., 2016].

Proposition 5. *Suppose that $(\mathcal{D}, \ell, \mathcal{F})$ satisfies the (B, β^*)-Bernstein condition for $\beta^* \in [0,1]$. Pick any $c > 0, \eta < 1/(2Bc)$. Then for all $0 < \beta \leq \beta^*$ and for all $f \in \mathcal{F}$:*

$$c \cdot \eta \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}\left[(\ell(f; Z') - \ell(f^*; Z'))^2\right] \leq \left(\frac{1}{2} \wedge \beta\right) \cdot \left(\mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(f; Z') - \ell(f^*; Z')]\right) + (1 - \beta) \cdot (2Bc\eta)^{\frac{1}{1-\beta}}$$

Note that, by our assumption on η, $\lim_{\beta \uparrow 1} (2Bc\eta)^{1/(1-\beta)} = 0$ and the second term vanishes for $\beta = 1$.

3 Main Development

Lemma 1 (Main technical lemma)**.** *Fix any two real numbers r_0, r_1 such that $|r_0|, |r_1| \leq 1$. Let $S \sim$ Ber(1/2) and let $\bar{S} = 1 - S$. Then for all $0 < \eta \leq 1/4$ it holds that*

$$r_{\bar{S}} - r_S \trianglelefteq_\eta \eta \cdot C_\eta r_{\bar{S}}^2 \leq \eta \cdot C_{1/4} r_{\bar{S}}^2$$

where $C_0 = 2$, C_η is a continuous increasing function of η and $C_{1/4} \approx 3.6064$.

The proof of this bound, with an explicit formula for the constant C_η, is in Appendix B. Our formula for C_η is tight near $\eta = 0$ but can be improved if it is known that r_0, r_1 are of the same sign. For ease of exposition, below we will only use the value for $\eta = 1/4$.

The lemma is the cornerstone in the proof of our main theorem which now follows. In this proof, r_S is set to the excess loss of a hypothesis f on an example from sample $\tilde{Z}_S$. Crucially, the square term on the right, when applied in the proof, only refers to a ghost sample $\tilde{Z}_{\bar{S}}$ while f is a hypothesis trained on the real sample $\tilde{Z}_{\mathbf{0}}$ – this allows us to 'kill' it under a Bernstein condition, replacing the square by a small enough linear term. A qualitatively similar inequality which has the sum $r_{\bar{S}}^2 + r_S^2$ on the right implicitly appears in [Audibert, 2004], but these square terms, being a combination of training and ghost samples, are not easily removed in our proof, and to get a PAC-Bayesian bound based on this lemma one needs to pick η small enough so that the term becomes negligible, leading to $\eta \asymp 1/\sqrt{n}$ which implies slow rates. Killing the square terms by taking a very small η also happens implicitly in the proof of the CMI result of Steinke and Zakynthinou [2020] as well as Hellström and Durisi [2020] which, for this reason, also give the slow rate.

We note that our Lemma 1 does not hold for unbounded losses and specifically does not hold for sub-Gaussian losses (to see this, for example, consider the case of $r_0 = 0$ and $r_1 < (-\ln 2)/\eta$). Adjusting this lemma for sub-Gaussian losses yields terms on the right-hand side that only lead to slow rates – a similar issue as the one described above occurring in prior work [Steinke and Zakynthinou, 2020, Hellström and Durisi, 2020, Audibert, 2004]. Thus, while a similar result as ours *might* hold for sub-Gaussian losses, it would require fundamentally new ideas to prove it.

Theorem 1. *Let $(\mathcal{D}, \ell, \mathcal{F})$ represent a learning problem which satisfies the (B, β^*)-Bernstein condition and suppose the loss function ℓ has range in $[0,1]$. Let $A : \bigcup_{i=1}^n \mathcal{Z}^i \to \Delta(\mathcal{F})$ be a possibly randomized learning algorithm and $\pi \in \Delta(\mathcal{F})$ be any almost exchangeable prior. Let $\tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}$ be two samples of n i.i.d. examples* *each drawn from $\mathcal{D}$. Then, for all $\beta \in [0, \beta^*]$, all $0 < \eta \leq \sqrt{n}\eta_{\max}/24$, it holds that,*

$$L(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) - L(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}}) \trianglelefteq_{\eta}^{\tilde{Z}_{\mathbf{0}}} (1 \wedge 2\beta) \cdot R(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}}) + 8 \cdot \left(\frac{\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}\rangle\right)\right] + \operatorname{llog} n}{n\eta_{\max}} \right)_{[**]}^{\frac{1}{2-\beta}} + \frac{6\eta}{n}, \tag{8}$$

where $\eta_{\max} = \left(\frac{1}{4} \wedge \frac{1}{2BC_{1/4}}\right)$*,* $C_{1/4} = 3.6064$*,* $\operatorname{llog} n = \log(\lceil \log_2(\sqrt{n}) \rceil + 2) = O(\log\log n)$ *and the notation* $a^b_{[**]}$ *stands for* $\max\{a^b, a\}$*.*

In all interesting cases, the quantity a inside the notation $a^b_{[**]}$ in the bound is less than 1, thus $a^b_{[**]} = a^b$. Otherwise, the bound would not be useful, as the LHS is less than 1 for any loss in $[0, 1]$.

Since this is still a $\trianglelefteq_\eta$-ESI statement with $\eta = \sqrt{n}\eta_{\max}/24$, it implies the in-probability statement that with probability at least $1 - \delta$, the above holds up to an additional $(-\log\delta)/\eta$ term on the right. More formally, a simple application of Proposition 1 to ESI (8) of Theorem 1 yields Corollary 1:

Corollary 1. *Consider the setting and notation of Theorem 1. Let $\delta \in (0, 1)$. For all $\beta \in [0, \beta^*]$ and all almost exchangeable priors π, with probability $1 - \delta$ over the choice of $\tilde{Z}_{\mathbf{0}} \sim \mathcal{D}^n$, we have*

$$L(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) - L(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}}) \leq (1 \wedge 2\beta) \cdot R(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}}) + 8 \cdot \left(\frac{\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}\rangle\right)\right] + \operatorname{llog} n}{n\eta_{\max}} \right)_{[**]}^{\frac{1}{2-\beta}} + \frac{\eta_{\max}}{4\sqrt{n}} + \frac{24\log(1/\delta)}{\sqrt{n}\eta_{\max}}.$$

The bound (8) also implies the corresponding in-expectation statement with the remainder term $6\eta/n$ set to 0. However, if one directly sets out to prove it, the term $\operatorname{llog} n$ and a factor of 2 from the multiplicative constant in front of the $a^b_{[**]}$ term can be avoided. In particular, the following improved bound holds, whose proof is based on the proof of Theorem 1 and is in Appendix B.

Corollary 2. (‘**Variation** of Theorem 1’) *Consider the setting and notation of Theorem 1. For all $\beta \in [0, \beta^*]$, it holds that*

$$\underset{\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[L(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) - L(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}})\right] \leq (1 \wedge 2\beta) \cdot \underset{\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[R(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}})\right] + 4 \cdot \left(\frac{\underset{\tilde{Z}_{\mathbf{0}},\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}\rangle\right)\right]}{n\eta_{\max}} \right)_{[**]}^{\frac{1}{2-\beta}}. \tag{9}$$

Moreover, for the right choice of prior, the expected KL term is $\mathsf{CMI}_{\mathcal{D}}(A)$, implying the bound:

Corollary 3. *Consider the setting and notation of Theorem 1. For all $\beta \in [0, \beta^*]$, it holds that*

$$\underset{\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[L(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) - L(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}})\right] \leq (1 \wedge 2\beta) \cdot \underset{\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[R(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}})\right] + 4 \cdot \left(\frac{\mathsf{CMI}_{\mathcal{D}}(A)}{n\eta_{\max}} \right)_{[**]}^{\frac{1}{2-\beta}}.$$

Proof of Corollary 3. Let $\tilde{Z} = (\tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}})$. We focus on the KL divergence in the bound (9):

$$\underset{\tilde{Z}_{\mathbf{0}},\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}\rangle\right)\right] = \underset{S,\tilde{Z}_{\mathbf{0}},\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}\rangle\right)\right] = \underset{S,\tilde{Z}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{S} \middle\| \pi|\langle \tilde{Z}_{S}, \tilde{Z}_{\bar{S}}\rangle\right)\right]$$

The first equality holds since S is independent of $\tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}$. The second equality holds because the distributions of $\tilde{Z}_S$, $\tilde{Z}_{\bar{S}}, \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}$ are all identical to $\mathcal{D}^n$ and π is almost exchangeable. We choose $\pi = \mathop{\mathbb{E}}_{S'}\left[A|\tilde{Z}_{S'}\right]$ for $S' \sim \mathrm{Ber}(1/2)^n$. Notice that π is indeed almost exchangeable. We now have

$$\mathop{\mathbb{E}}_{S,\tilde{Z}}\left[\mathsf{KL}\left(A|\tilde{Z}_S \middle\| \mathop{\mathbb{E}}_{S'}\left[A|\tilde{Z}_{S'}\right]\right)\right] = \mathop{\mathbb{E}}_{\tilde{Z}}\left[\mathop{\mathbb{E}}_{S}\left[\mathsf{KL}\left(A|\tilde{Z}_S \middle\| \mathop{\mathbb{E}}_{S'}\left[A|\tilde{Z}_{S'}\right]\right)\right]\right] = \mathop{\mathbb{E}}_{\tilde{Z}}\left[I^{\tilde{Z}}(A|\tilde{Z}_S; S)\right] = \mathsf{CMI}_{\mathcal{D}}(A).$$

Combining the two equations and substituting the term in inequality (9) completes the proof. □

After observing the implications of Theorem 1, we now present its complete proof below.

Proof of Theorem 1. Let $\tilde{z} = ((\tilde{z}_{1,0}, \tilde{z}_{1,1}), \ldots, (\tilde{z}_{n,0}, \tilde{z}_{n,1}))^\top \in \mathcal{Z}^{n\times 2}$ be a given, fixed supersample. Let $S = (S_1, \ldots, S_n)$, with $S_1, S_2, \ldots, S_n$ i.i.d. Ber$(1/2)$, be a selection vector and let $\bar{S}$ be its complement, that is, $\bar{S}_i := 1 - S_i$ for all $i \in [n]$. For each fixed $f \in \mathcal{F}$ and $\tilde{z} \in \mathcal{Z}^{n\times 2}$, we define

$$r_i(f; \tilde{z}_{i,0}) = \ell(f; \tilde{z}_{i,0}) - \ell(f^*; \tilde{z}_{i,0}) \;\text{ and }\; r_i(f; \tilde{z}_{i,1}) = \ell(f; \tilde{z}_{i,1}) - \ell(f^*; \tilde{z}_{i,1}).$$

Since ℓ has range in $[0, 1]$, it holds that for all $i \in [n]$, $|r_i(f; \tilde{z}_{i,0})|, |r_i(f; \tilde{z}_{i,1})| \leq 1$. By Lemma 1, for all $i \in [n]$, and $\eta < 1/4$, it holds that

$$r_i(f; \tilde{z}_{i,\bar{S}_i}) - r_i(f; \tilde{z}_{i,S_i}) \trianglelefteq_\eta^{S_i} \eta C_{1/4} r_i^2(f, \tilde{z}_{i,\bar{S}_i}) \tag{10}$$

Now take randomness under the product distribution Ber$(1/2)^n$ of S. By independence of the S_i and applying Proposition 2, we can then add the n ESIs (10) to give:

$$\sum_{i=1}^n r_i(f; \tilde{z}_{i,\bar{S}_i}) - \sum_{i=1}^n r_i(f; \tilde{z}_{i,S_i}) \trianglelefteq_\eta^{S} \eta C_{1/4} \sum_{i=1}^n r_i^2(f, \tilde{z}_{i,\bar{S}_i}).$$

Now consider a learning algorithm A that outputs a distribution $A|\tilde{z}_S$ on $\mathcal{F}$, and any 'prior' distribution $\pi|\tilde{z}$ on $\mathcal{F}$ that is allowed to depend on $\tilde{z}$ (which for now is considered fixed). The PAC-Bayes theorem (Proposition 3) gives

$$\mathop{\mathbb{E}}_{f\sim A|\tilde{z}_S}\left[\sum_{i=1}^n r_i(f; \tilde{z}_{i,\bar{S}_i}) - \sum_{i=1}^n r_i(f; \tilde{z}_{i,S_i})\right] \trianglelefteq_\eta^{S|\tilde{z}} \eta C_{1/4} \mathop{\mathbb{E}}_{f\sim A|\tilde{z}_S}\left[\sum_{i=1}^n r_i^2(f, \tilde{z}_{i,\bar{S}_i})\right] + \frac{\mathsf{KL}(A|\tilde{z}_S \| \pi|\tilde{z})}{\eta}. \tag{11}$$

We note that S is independent of $\tilde{z}$, so the ESI above could be equivalently written with respect to S instead of $S|\tilde{z}$.

Since inequality (11) holds for *all* $\tilde{z}$, we weaken it to an ESI by taking its expectation over $\tilde{Z} \sim \mathcal{D}^{n\times 2}$:

$$\mathop{\mathbb{E}}_{f\sim A|\tilde{Z}_S}\left[\sum_{i=1}^n r_i(f; \tilde{Z}_{i,\bar{S}_i}) - \sum_{i=1}^n r_i(f; \tilde{Z}_{i,S_i})\right] \trianglelefteq_\eta^{S,\tilde{Z}} \eta C_{1/4} \mathop{\mathbb{E}}_{f\sim A|\tilde{Z}_S}\left[\sum_{i=1}^n r_i^2(f, \tilde{Z}_{i,\bar{S}_i})\right] + \frac{\mathsf{KL}\left(A|\tilde{Z}_S \middle\| \pi|\tilde{Z}\right)}{\eta}$$

Since the conditional distribution π is almost exchangeable with respect to $\tilde{z}$, the above is rewritten as

$$\mathop{\mathbb{E}}_{f\sim A|\tilde{Z}_S}\left[\sum_{i=1}^n r_i(f; \tilde{Z}_{i,\bar{S}_i}) - \sum_{i=1}^n r_i(f; \tilde{Z}_{i,S_i})\right] \trianglelefteq_\eta^{S,\tilde{Z}} \eta C_{1/4} \mathop{\mathbb{E}}_{f\sim A|\tilde{Z}_S}\left[\sum_{i=1}^n r_i^2(f, \tilde{Z}_{i,\bar{S}_i})\right] + \frac{\mathsf{KL}\left(A|\tilde{Z}_S \middle\| \pi|(\tilde{Z}_S, \tilde{Z}_{\bar{S}})\right)}{\eta}.$$

Now, since the $\tilde{Z}_{1,0}, \tilde{Z}_{1,1}, \ldots, \tilde{Z}_{n,0}, \tilde{Z}_{n,1}$ are i.i.d. and independent of the S_i, we must also have:

$$\mathop{\mathbb{E}}_{f\sim A|\tilde{Z}_{\mathbf{0}}}\left[\sum_{i=1}^n r_i(f; \tilde{Z}_{i,1}) - \sum_{i=1}^n r_i(f; \tilde{Z}_{i,0})\right] \trianglelefteq_\eta^{\tilde{Z}} \eta C_{1/4} \mathop{\mathbb{E}}_{f\sim A|\tilde{Z}_{\mathbf{0}}}\left[\sum_{i=1}^n r_i^2(f, \tilde{Z}_{i,1})\right] + \frac{\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\langle\tilde{Z}\rangle\right)}{\eta},$$

where we also replaced $\pi|\tilde{z}$ by its equivalent $\pi|\langle\tilde{z}\rangle$. Since the $\tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}}$ consist of i.i.d. random variables, we can weaken the above inequality to an in-expectation inequality (by Proposition 1) with respect to the 'ghost" sample $\tilde{Z}_{\mathbf{1}} \sim \mathcal{D}^n$:

$$\underset{f\sim A|\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\sum_{i=1}^{n} r_i(f;\tilde{Z}_{i,1}) - \sum_{i=1}^{n} r_i(f;\tilde{Z}_{i,0})\right]\right] \trianglelefteq_{\eta}^{\tilde{Z}_{\mathbf{0}}}$$

$$\eta C_{1/4} \underset{f\sim A|\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\sum_{i=1}^{n} r_i^2(f,\tilde{Z}_{i,1})\right]\right] + \underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\frac{\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}}\middle\|\pi|\langle\tilde{Z}\rangle\right)}{\eta}\right]. \quad (12)$$

We now focus on term of the expected sum of squared excess risks in the RHS. By applying the linearized (B, β^*)-Bernstein condition of Proposition 5 and adding the inequalities for all $i \in [n]$, we have that for all $\eta < 1/(2BC_{1/4})$, $\beta \in [0, \beta^*]$,

$$\eta C_{1/4} \underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\sum_{i=1}^{n} r_i^2(f,\tilde{Z}_{i,1})\right] \leq \left(\frac{1}{2}\wedge\beta\right)\cdot \underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\sum_{i=1}^{n} r_i(f;\tilde{Z}_{i,1})\right] + n(1-\beta)(2BC_{1/4}\eta)^{1/(1-\beta)}. \quad (13)$$

Now, observe that $\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\sum_{i=1}^{n} r_i(f;\tilde{Z}_{i,1})\right] = n\cdot R(f;\mathcal{D})$ and $\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\sum_{i=1}^{n} r_i(f;\tilde{Z}_{i,0})\right] = n\cdot R(f;\tilde{Z}_{\mathbf{0}})$. Combining inequality (12) with (13) and substituting the terms above, we have that for all $\eta < \eta_{\max} := \left(\frac{1}{4}\wedge\frac{1}{2BC_{1/4}}\right)$,

$$\underset{f\sim A|\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}\left[n\cdot R(f;\mathcal{D}) - n\cdot R(f;\tilde{Z}_{\mathbf{0}})\right] \trianglelefteq_{\eta}^{\tilde{Z}_{\mathbf{0}}}$$

$$\left(\frac{1}{2}\wedge\beta\right)\cdot \underset{f\sim A|\tilde{Z}_{\mathbf{0}}}{\mathbb{E}}[n\cdot R(f;\mathcal{D})] + n(1-\beta)(2BC_{1/4}\eta)^{1/(1-\beta)} + \underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\frac{\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}}\middle\|\pi|\langle\tilde{Z}\rangle\right)}{\eta}\right].$$

Dividing by n and substituting for the expected true and empirical excess risk of the randomized estimator $A|\tilde{Z}_{\mathbf{0}}$, we have the following ESI:

$$R(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) \;-\; R(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}}) \;\trianglelefteq_{n\eta}^{\tilde{Z}_{\mathbf{0}}}\; \left(\frac{1}{2}\wedge\beta\right)\cdot R(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) + \left(\frac{\eta}{\eta_{\max}}\right)^{\frac{1}{1-\beta}} + \frac{\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}}\middle\|\pi|\langle\tilde{Z}\rangle\right)\right]}{n\eta}. \quad (14)$$

Using Proposition 4, we now extend this ESI to deal with random η. The proposition immediately gives that for every finite grid $\mathcal{G} \subset [\eta_{\min}, \eta_{\max}]$, for arbitrary probability mass function $\pi_{\mathcal{G}}$ on $\mathcal{G}$, for arbitrary functions (random variables) $\hat{\eta} : \tilde{Z}_{\mathbf{0}} \to \mathcal{G}$, we have:

$$R(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) \;-\; R(A|\tilde{Z}_{\mathbf{0}};\tilde{Z}_{\mathbf{0}}) \;\trianglelefteq_{n\eta_{\min}}^{\tilde{Z}_{\mathbf{0}}}\; \left(\frac{1}{2}\wedge\beta\right)\cdot R(A|\tilde{Z}_{\mathbf{0}};\mathcal{D}) + \left(\frac{\hat{\eta}}{\eta_{\max}}\right)^{\frac{1}{1-\beta}} + \frac{\textsc{ub} - \log\pi_{\mathcal{G}}(\hat{\eta})}{n\hat{\eta}}, \quad (15)$$

where $\textsc{ub}$ can be any upper bound on $\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}}\middle\|\pi|\langle\tilde{Z}\rangle\right)\right]$. In the remainder of the proof we simply set $\textsc{ub} = \mathbb{E}_{\tilde{Z}_{\mathbf{1}}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}}\middle\|\pi|\langle\tilde{Z}\rangle\right)\right]$, the possibility to take a larger upper bound is explored in Example 2.

Now let $\pi_{\mathcal{G}}$ be the uniform distribution over the grid

$$\mathcal{G} := \left\{\eta_{\max}, \frac{1}{2}\eta_{\max}, \ldots, \frac{1}{2^K}\eta_{\max} : K := \lceil\log_2(\sqrt{n})\rceil + 2\right\} \quad (16)$$

and define $\hat{\eta}'$, as function of data $\tilde{Z}_{\mathbf{0}}$ to be the element of $[0, \eta_{\max}]$ minimizing the sum

$$\textsc{comp}(\eta) = \left(\frac{\eta}{\eta_{\max}}\right)^{\frac{1}{1-\beta}} + \frac{\underset{\tilde{Z}_{\mathbf{1}}}{\mathbb{E}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}}\middle\|\pi|\tilde{Z}\right)\right] - \log\pi_{\mathcal{G}}(\eta)}{n\eta}$$

of the last two terms in (15), and let $\hat{\eta}$ be the element within $\mathcal{G}$ that minimizes this sum. We can determine $\hat{\eta}'$ by differentiation. We find that, since we have $|\mathcal{G}| = K+1 \geq 3$ and hence $-\log \pi_{\mathcal{G}}(\hat{\eta}) \geq 1$, it holds

$$\text{COMP}(\hat{\eta}) \leq \begin{cases} 2 \cdot \text{COMP}(\hat{\eta}') = 4 \left(\frac{\mathbb{E}_{\tilde{Z}_{\mathbf{1}}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\tilde{Z}\right)\right] + \text{llog } n}{n\eta_{\max}} \right)^{1/(2-\beta)} & \text{if } \hat{\eta}' < \eta_{\max} \\ \text{COMP}(\hat{\eta}') \leq 2 \left(\frac{\mathbb{E}_{\tilde{Z}_{\mathbf{1}}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\tilde{Z}\right)\right] + \text{llog } n}{n\eta_{\max}} \right) & \text{if } \hat{\eta}' = \eta_{\max} \end{cases}$$

where $\text{llog } n = \log(\lceil \log_2(\sqrt{n}) \rceil + 2) = O(\log \log n)$. Combining this with (15) gives

$$R(A|\tilde{Z}_{\mathbf{0}}; \mathcal{D}) \; - \; R(A|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}}) \; \trianglelefteq^{\tilde{Z}_{\mathbf{0}}}_{n\eta_{\min}} \; \alpha \cdot R(A|\tilde{Z}_{\mathbf{0}}; \mathcal{D}) \, + \, 4 \cdot \left(\frac{\mathbb{E}_{\tilde{Z}_{\mathbf{1}}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\tilde{Z}\right)\right] + \text{llog } n}{n\eta_{\max}} \right)^{1/(2-\beta)}_{[**]} \tag{17}$$

for every $0 < \eta_{\min} \leq \frac{\eta_{\max}}{8\sqrt{n}}$, since we have:

$$\hat{\eta} \geq \frac{\eta_{\max}}{2^K} = \frac{\eta_{\max}}{2^{\lceil \log_2(\sqrt{n}) \rceil + 2}} \geq \frac{\eta_{\max}}{2^{\log_2(\sqrt{n}) + 3}} = \frac{\eta_{\max}}{8\sqrt{n}}.$$

Here the notation $a^b_{[**]}$ indicates $\max\{a^b, a\}$ and here and below we set $\alpha = \left(\frac{1}{2} \wedge \beta\right)$.

From inequality (17), we can derive the following two ESIs. First, by substituting $R(A|\tilde{Z}_{\mathbf{0}}; \mathcal{D})$ and $R(A|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}})$ and $\eta := n\eta_{\min}$ and rearranging, we have for every $\eta \leq \sqrt{n}\eta_{\max}/8$ that

$$\begin{aligned} &L(A|\tilde{Z}_{\mathbf{0}}; \mathcal{D}) - L(A|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}}) \trianglelefteq^{\tilde{Z}_{\mathbf{0}}}_{\eta} \\ &\qquad \alpha \cdot R(A|\tilde{Z}_{\mathbf{0}}; \mathcal{D}) + 4 \cdot \left(\frac{\mathbb{E}_{\tilde{Z}_{\mathbf{1}}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\tilde{Z}\right)\right] + \text{llog } n}{n\eta_{\max}} \right)^{1/(2-\beta)}_{[**]} + L(f^*; \mathcal{D}) - L(f^*; \tilde{Z}_{\mathbf{0}}) \end{aligned} \tag{18}$$

Second, by rearranging and multiplying by $\alpha/(1-\alpha)$, (17) also gives

$$\alpha R(A|\tilde{Z}_{\mathbf{0}}; \mathcal{D}) \trianglelefteq^{\tilde{Z}_{\mathbf{0}}}_{\eta(1-\alpha)/\alpha} 2\alpha \cdot \left(R(A|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}}) + 4 \cdot \left(\frac{\mathbb{E}_{\tilde{Z}_{\mathbf{1}}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\tilde{Z}\right)\right] + \text{llog } n}{n\eta_{\max}} \right)^{1/(2-\beta)}_{[**]} \right), \tag{19}$$

where we used that $\alpha \leq 1/2$ hence $\alpha/(1-\alpha) \leq 1$ and the fact that, straightforwardly, $U \trianglelefteq_{\eta} 0 \Rightarrow cU \trianglelefteq_{\eta/c} 0$. We want to combine these two ESIs, while also replacing the final term $L(f^*; \mathcal{D}) - L(f^*; \tilde{Z}_{\mathbf{0}})$ in (18). For this we note that Hoeffding's Lemma in ESI notation combined with the ESI chain rule (Proposition 2) for i.i.d. random variables immediately gives $n(L(f^*; \mathcal{D}) - L(f^*; \tilde{Z}_{\mathbf{0}})) \trianglelefteq_{\eta'} 2n\eta'$ for all $\eta' > 0$, hence also $L(f^*; \mathcal{D}) - L(f^*; \tilde{Z}_{\mathbf{0}}) \trianglelefteq_{n\eta'} 2\eta'$ and hence substituting $\eta := \eta' n$,

$$L(f^*; \mathcal{D}) - L(f^*; \tilde{Z}_{\mathbf{0}}) \trianglelefteq_{\eta} \frac{2\eta}{n}. \tag{20}$$

Chaining ESIs (18), (19) and (20), using Proposition 2(a), now gives, for all $\eta \leq \sqrt{n}\eta_{\max}/8$,

$$\begin{aligned} &L(A|\tilde{Z}_{\mathbf{0}}; \mathcal{D}) - L(A|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}}) \trianglelefteq^{\tilde{Z}_{\mathbf{0}}}_{\eta(1-\alpha)/(2-\alpha)} \\ &\qquad (1 \wedge 2\beta) \cdot R(A|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}}) + 8 \cdot \left(\frac{\mathbb{E}_{\tilde{Z}_{\mathbf{1}}}\left[\mathsf{KL}\left(A|\tilde{Z}_{\mathbf{0}} \middle\| \pi|\tilde{Z}\right)\right] + \text{llog } n}{n\eta_{\max}} \right)^{1/(2-\beta)}_{[**]} + \frac{2\eta}{n}. \end{aligned} \tag{21}$$

Since, by $0 \leq \alpha \leq 1/2$, $(1-\alpha)/(2-\alpha) \geq 1/3$, the result follows substituting η in place of $\eta/3$. □

3.1 Applications

In this section, we demonstrate some applications of Theorem 1, providing classes $\mathcal{F}$ for which standard PAC-Bayesian bounds are suboptimal or difficult to obtain, but the almost exchangeable priors conditioned on supersamples make them straightforward. We note that the settings are slight extensions of examples already covered by Audibert [2004] and Steinke and Zakynthinou [2020] in the non-fast rate setting; the added benefit is the fast-rate treatment allowed by Theorem 1 and its extension for Gibbs posteriors in Example 2. For starters, the following observation (proof omitted) allows us to mix almost exchangeable priors and to construct them from standard priors:

Proposition 6. *Let W be any standard distribution on $\mathcal{F}$ independent of the data, i.e. $W|\langle\tilde{z}\rangle = W|\langle\tilde{z}'\rangle$ for all $\tilde{z}, \tilde{z}' \in \mathcal{Z}^{2n}$. Then W is also an almost exchangeable prior. Further, let $\{W_k : k \in \mathbb{N}\}$ denote a countable set of almost exchangeable priors and let ρ be a probability mass function on $\mathbb{N}$. Then W defined by $W \mid \langle\tilde{z}\rangle = \sum_{k\in\mathbb{N}} \rho(k) \cdot W_k \mid \langle\tilde{z}\rangle$ is an almost exchangeable prior as well.*

3.1.1 VC classes

In this section, $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ and $\mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$ is a hypothesis class with VC dimension d. We work with the 0-1 loss $\ell : \mathcal{F} \times (\mathcal{X} \times \{0, 1\}) \to \{0, 1\}$ defined by $\ell(f,(x, y)) = 0 \Leftrightarrow f(x) = y$.

Theorem 2. *Let $\mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$ be a hypothesis class with VC dimension d and let $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$. There exists a deterministic Empirical Risk Minimization algorithm $A : \mathcal{Z}^* \to \mathcal{F}$ for 0/1 loss and an almost exchangeable prior π, such that, for any $\tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}} \in \mathcal{Z}^n$,*

$$\mathsf{KL}(A|\tilde{z}_{\mathbf{0}} \| \pi|\langle\tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}}\rangle) \leq d \log(2n).$$

The theorem can be proven by following the same steps as the proof of Steinke and Zakynthinou [2020, Theorem 4.12]; we provide its proof in Appendix B.

Example 1 (Thresholds)**.** Consider the set of threshold functions $\mathcal{T} = \{f_t : \mathbb{N} \to \{0, 1\} : t \in \mathbb{N} \cup \{\infty\}\}$, where $f_t(x) = 1 \Leftrightarrow x \geq t$. Let ℓ be the 0/1 loss satisfying $\ell(f,(x, y)) = 0 \Leftrightarrow f(x) = y$. Let $A : (\mathbb{N} \times \{0, 1\})^n \to \mathcal{T}$ be a learning algorithm that outputs the smallest optimal threshold – i.e., $A|z = f_{\min\{x : f_x \in \arg\min_{f\in\mathcal{T}} \ell(f,z)\}\cup\{\infty\}}$. It is straightforward to see that A is an ERM that satisfies the global consistency property from the proof of Theorem 2. Since $\mathcal{T}$ has VC dimension $d = 1$, there exists an almost exchangable prior π such that $\mathsf{KL}(A|\tilde{z}_{\mathbf{0}} \| \pi|\langle\tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}}\rangle) \leq \log(2n)$ for all $\tilde{z}$. Now suppose that we have a distribution $\mathcal{D}$ with random label noise – i.e., there is some t^* such that each data point X_i is sampled from an arbitrary $\mathcal{D}_{\mathcal{X}}$ and, given X_i, $Y_i = f_{t^*}(X_i)$ with probability $1-p$ and $Y_i = 1-f_{t^*}(X_i)$ with probability p for some $0 < p < 1/2$. This implies the Massart condition and hence the Bernstein condition with $\beta = 1$ and B depending on p [Van Erven et al., 2015]. Still, the empirical error of ERM does not go to 0 with n due to the label noise. Therefore, standard PAC-Bayes bounds (1) are of order $\sqrt{\mathsf{KL}/n}$, whereas Theorem 1 gives a fast rate of order $(\log n)/n$.

3.1.2 Compression Scheme Priors

The following extends the notion of a compression scheme due to Littlestone and Warmuth [1986].

Definition 9 (Compression Scheme Prior)**.** *We call a data-dependent distribution $W : \mathcal{Z}^n \to \Delta(\mathcal{F})$ a* compression scheme prior of size k *if we can write $W|z = W_2|(A_1|z)$ for all z, where*

1. *$A_1 : \mathcal{Z}^n \to \mathcal{Z}^k$ is a "compression algorithm" which given a sample $z \in \mathcal{Z}^n$ selects a subset $i_1, \ldots, i_k \in [n]$ and returns $(z_{i_1}, \ldots, z_{i_k}) \in \mathcal{Z}^k$ and*
2. *$W_2 : \mathcal{Z}^k \to \Delta(\mathcal{F})$ is any function.*

For $k = 0$, we say that W is a compression scheme prior of size 0 iff it outputs a fixed distribution.

Theorem 3. *Let $W : \mathcal{Z}^n \to \Delta(\mathcal{F})$ be a compression scheme prior of size $k \geq 0$ and $A : \mathcal{Z}^n \to \Delta(\mathcal{F})$ be an arbitrary possibly randomized learning algorithm. Then there exists an almost exchangeable prior π, such that for all $\tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}} \in \mathcal{Z}^n$,*

$$\mathsf{KL}(A|\tilde{z}_{\mathbf{0}} \| \pi|\langle\tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}}\rangle) \leq \mathsf{KL}(A|\tilde{z}_{\mathbf{0}} \| W|\tilde{z}_{\mathbf{0}}) + k \log(2n). \tag{22}$$

Proof of Theorem 3. Let $W = W_2|(A_1|z)$ be a compression scheme prior and let $\langle \tilde z \rangle = \langle \tilde z_{\mathbf 0}, \tilde z_{\mathbf 1} \rangle$. We choose the conditional prior distribution as

$$\pi(f|\langle \tilde z \rangle) = \frac{\sum_{z^k \in K(\tilde z)} \mathcal{P}_{W_2|z^k}(f)}{\binom{2n}{k}},$$

where we denote by $K(z)$ the set of all subsets of z of size k. Observe that π is indeed an almost exchangeable prior. It holds that

$$\begin{aligned}
\mathsf{KL}(A|\tilde z_{\mathbf 0} \| \pi|\langle \tilde z_{\mathbf 0}, \tilde z_{\mathbf 1}\rangle) &= \mathop{\mathbb{E}}_{f \sim A|\tilde z_{\mathbf 0}} \left[\log \frac{\mathcal{P}_{A|\tilde z_{\mathbf 0}}(f)}{\pi(f|\langle \tilde z \rangle)} \right] \\
&= \mathop{\mathbb{E}}_{f \sim A|\tilde z_{\mathbf 0}} \left[\log \frac{\mathcal{P}_{A|\tilde z_{\mathbf 0}}(f) \cdot \binom{2n}{k}}{\sum_{z^k \in K(\tilde z)} \mathcal{P}_{W_2|z^k}(f)} \right] \\
&\leq \mathop{\mathbb{E}}_{f \sim A|\tilde z_{\mathbf 0}} \left[\log \frac{\mathcal{P}_{A|\tilde z_{\mathbf 0}}(f) \cdot \binom{2n}{k}}{\mathcal{P}_{W_2|(A_1|\tilde z_{\mathbf 0})}(f)} \right] \\
&= \mathop{\mathbb{E}}_{f \sim A|\tilde z_{\mathbf 0}} \left[\log \frac{\mathcal{P}_{A|\tilde z_{\mathbf 0}}(f)}{\mathcal{P}_{W|\tilde z_{\mathbf 0}}(f)} \right] + \log \binom{2n}{k} \\
&\leq \mathsf{KL}(A|\tilde z_{\mathbf 0} \| W|\tilde z_{\mathbf 0}) + k \log(2n),
\end{aligned}$$

where the first inequality holds since $A_1|\tilde z_{\mathbf 0} \in K(\tilde z)$ which implies that $\sum_{z^k \in K(\tilde z)} \mathcal{P}_{W_2|z^k}(f) \geq \mathcal{P}_{W_2|(A_1|\tilde z_{\mathbf 0})}(f)$. The last inequality follows by the common bound $\binom{2n}{k} \leq (2n)^k$. □

In the case that we choose a size k compression scheme prior W that, upon each input, puts all its mass on a single $\hat f \mid \tilde Z_{\mathbf 0} \in \mathcal{F}$, and we choose A to be the deterministic learning algorithm that is *equal* to W, then A has a compression scheme of size k in the original sense of Littlestone and Warmuth [1986] and its KL complexity will by Theorem 3 be bounded by $k \log(2n)$. Our generalization allows us to choose an algorithm A different from W that might, for example, base its output on the whole dataset and not just the k points selected by A_1 'inside' W. An example algorithm with pleasant properties is the *Gibbs algorithm* with W as a prior.

Example 2. (Gibbs Algorithm based on Compression Scheme Prior) The *Gibbs* or *generalized Bayes* learning algorithm (see, e.g., Alquier [2020], Grünwald and Mehta [2020], Zhang [2006b]) $A_{\text{Gibbs}} : \mathcal{Z}^n \to \Delta(\mathcal{F})$ with (possibly data-dependent) learning rate $\hat\eta$ based on data-dependent prior distribution W is defined in terms of its posterior density (Radon-Nikodym derivative) relative to W, as

$$\frac{d(A_{\text{Gibbs}}|\tilde Z_{\mathbf 0})}{d(W|\tilde Z_{\mathbf 0})}(f) \propto \exp(-\hat\eta n R(f; \tilde Z_{\mathbf 0})).$$

This is the standard definition of the Gibbs algorithm relative to prior distribution $W \mid \tilde Z_{\mathbf 0}$. A modification of the proof of Theorem 1, sketched in Appendix B, gives the following corollary: *if we set A to the Gibbs algorithm relative to size k compression scheme prior W, and A' to any other algorithm, we have, with the same abbreviations as in Theorem 1,*

$$\begin{aligned}
&L(A_{\text{Gibbs}}|\tilde Z_{\mathbf 0}; \mathcal{D}) \trianglelefteq_{\eta}^{\tilde Z_{\mathbf 0}} \\
&\qquad L(A'|\tilde Z_{\mathbf 0}; \tilde Z_{\mathbf 0}) + (1 \wedge 2\beta) \cdot R(A'|\tilde Z_{\mathbf 0}; \tilde Z_{\mathbf 0}) + 8 \cdot \left(\frac{\mathsf{KL}\left(A'|\tilde Z_{\mathbf 0} \middle\| W|\tilde Z_{\mathbf 0}\right) + O(k \log n)}{n \eta_{\max}} \right)^{1/(2-\beta)}_{[**]} + \frac{6\eta}{n}.
\end{aligned}$$

In particular, if A' is set to an ERM, the sum of the first two terms on the right is upper bounded by $L(A_{\text{Gibbs}}|\tilde Z_{\mathbf 0}; \tilde Z_{\mathbf 0})$ again and, under a Bernstein condition, we get a fast rate for the Gibbs algorithm as well, although the complexity term is taken relative to ERM rather than Gibbs.

4 Conclusion and Future Work

We have shown how to extend PAC-Bayesian and Mutual Information Bounds to a fast-rate conditional version which allows us to handle arbitrary VC classes. One point which remains open for future research is the fact that, unless we use ERM and we deal with losses like the squared error for which we know the β for which the Bernstein condition holds in advance, the bound (6) is not empirical (observable from data only, without knowing $\mathcal{D}$ or f^*). Mhammedi et al. [2019] do provide an empirically observable bound that achieves fast rates, by replacing f^* by an estimator based on part of the training data only (a technique called *de-biasing* by Y. Seldin) and by replacing the $O((\mathsf{KL}/n)^{1/(2-\beta)})$ term by an empirical variance-like term that goes to 0 at the right rate if a Bernstein condition holds but can be calculated without knowing β. It seems likely that our bound can also be made fully empirical, for arbitrary learning algorithms and losses rather than just ERM and curved losses. Whether this is really the case will be sorted out in future work. Another interesting open question is whether a similar bound holds for unbounded but sub-Gaussian losses; see the discussion underneath Lemma 1.

Acknowledgements

We thank the reviewers of COLT for useful comments on the presentation of our manuscript. LZ was supported by a Facebook Fellowship. PG was supported by the Dutch Research Council (NWO) via research programme 617.001.651. This work began when TS and LZ were at IBM Research in Almaden.

References

Pierre Alquier. Approximate Bayesian inference. *Entropy*, 22(11), 2020. ISSN 1099-4300. doi: 10.3390/e22111272. URL `https://www.mdpi.com/1099-4300/22/11/1272`.

Amiran Ambroladze, Emilio Parrado-hernández, and John Shawe-taylor. Tighter PAC-Bayes bounds. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 9–16. MIT Press, 2007. URL `https://proceedings.neurips.cc/paper/2006/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

J.Y. Audibert. *PAC-Bayesian statistical learning theory.* PhD thesis, Université Paris VI, 2004.

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135 (3):311–334, 2006.

Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 25–55. PMLR, 07–09 Apr 2018. URL `http://proceedings.mlr.press/v83/bassily18a.html`.

Olivier Catoni. *PAC-Bayesian Supervised Classification.* Lecture Notes-Monograph Series. IMS, 2007.

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

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in pac-bayes. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 604–612. PMLR, 13–15 Apr 2021. URL `http://proceedings.mlr.press/v130/karolina-dziugaite21a.html`.

Tim van Erven, P. Grünwald, N. Mehta, M. Reid, and R. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 2015. URL http://arxiv.org/abs/1502.1507.02592. Special issue in Memory of Alexey Chervonenkis.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20, 2015.

Peter D. Grünwald and Nishant A. Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Proceedings of the Thirtieth Conference on Algorithmic Learning Theory (ALT) 2019*, 2019.

Peter D. Grünwald and Nishant A. Mehta. Fast rates for general unbounded loss functions: From ERM to generalized Bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020. URL http://jmlr.org/papers/v21/18-488.html.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *arXiv preprint arXiv:2004.12983*, 2020.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *arXiv preprint arXiv:2005.08044*, 2020.

Wouter M Koolen, Peter Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 4457–4465. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/db116b39f7a3ac5366079b1d9fe249a5-Paper.pdf.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Technical report, 1986.

Roi Livni and Shay Moran. A limitation of the PAC-Bayes framework. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper/2020/file/ec79d4bed810ed64267d169b0d37373e-Paper.pdf.

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

D. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh ACM Conference on Computational Learning Theory (COLT' 98)*, pages 230–234. ACM Press, 1998.

D. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected Bernstein inequality. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 12202–12213. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3dea6b598a16b334a53145e78701fa87-Paper.pdf.

Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. A direct sum result for the information complexity of learning. *arXiv preprint arXiv:1804.05474*, 2018.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pages 11013–11023, 2019.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/russo16.html.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145 – 147, 1972. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(72)90019-2. URL http://www.sciencedirect.com/science/article/pii/0097316572900192.

M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

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

Thomas Steinke and Lydia Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/steinke20a.html.

Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems*, pages 109–117, 2013.

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, January 1971. doi: 10.1137/1116025.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

Jun Yang, Shengyang Sun, and Daniel M Roy. Fast-rate pac-bayes generalization bounds via shifted rademacher processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/9715d04413f296eaf3c30c47cec3daa6-Paper.pdf.

Ernst Zermelo. Beweis, daß jede menge wohlgeordnet werden kann. *Mathematische Annalen*, 59(4):514–516, 1904.

Tong Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006a.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *ICLR*, 2019.

A Glossary

Notation	Description
$\mathcal{D}$	Probability distribution over $\mathcal{Z}$
Z	i.i.d. sample of size n: $Z = (Z_1, \dots, Z_n) \sim \mathcal{D}^n$
$(\mathcal{D}, \ell, \mathcal{F})$	Learning problem for distribution $\mathcal{D}$, loss function ℓ, and set of hypotheses $\mathcal{F}$
$\ell(f; z)$	Empirical loss of f on sample $z \in \mathcal{Z}^n$: $\ell(f; z) = \frac{1}{n} \sum_{i=1}^{n} \ell(f; z_i)$
$\ell(f; \mathcal{D})$	True loss of f: $\mathbb{E}_{Z' \sim \mathcal{D}}[\ell(f; Z')]$
f^*	True loss minimizer within $\mathcal{F}$: $\ell(f^*; \mathcal{D}) = \inf_{f \in \mathcal{F}} \ell(f; \mathcal{D})$
A	(Possibly randomized) learning algorithm: $A : \bigcup_{i=1}^{n} \mathcal{Z}^i \to \Delta(\mathcal{F})$
$A \vert Z$	Posterior distribution of output of A on input $Z \sim \mathcal{D}^n$
$L(F; z)$	Empirical loss of $F \in \Delta(\mathcal{F})$ on sample $z \in \mathcal{Z}^n$: $L(F; z) = \mathbb{E}_{f \sim F}[\ell(f; z)]$
$L(F; \mathcal{D})$	True loss of $F \in \Delta(\mathcal{F})$: $L(F; \mathcal{D}) = \mathbb{E}_{f \sim F}[\ell(f; \mathcal{D})]$
$R(F; z)$	Empirical excess risk of $F \in \Delta(\mathcal{F})$ on sample $z \in \mathcal{Z}^n$: $R(F; z) = \mathbb{E}_{f \sim A}[r(f; z)]$
$R(F; \mathcal{D})$	True excess risk of $F \in \Delta(\mathcal{F})$: $R(F; \mathcal{D}) = \mathop{\mathbb{E}}_{f \sim A}[r(f; \mathcal{D})]$
$\tilde{Z}$	Supersample $\tilde{Z} = \left((\tilde{Z}_{1,0}, \tilde{Z}_{1,1}), \dots, (\tilde{Z}_{n,0}, \tilde{Z}_{n,1})\right)^\top \sim \mathcal{D}^{n \times 2}$
S	Random selector vector $S \sim \mathrm{Ber}(1/2)^n$
$\tilde{Z}_S$	Subset of $\tilde{Z}$ indexed by $S \in \{0,1\}^n$: $\tilde{Z}_S = (Z_{1,S_1}, \dots, Z_{n,S_n})^\top \in \mathcal{Z}^n$
$\langle \tilde{Z} \rangle$	List of *unordered* pairs of $\tilde{Z}$: $\langle \tilde{Z} \rangle = \langle \tilde{Z}_{\mathbf{0}}, \tilde{Z}_{\mathbf{1}} \rangle = (\{\tilde{Z}_{1,0}, \tilde{Z}_{1,1}\}, \dots, \{\tilde{Z}_{n,0}, \tilde{Z}_{n,1}\})^\top$

Table 2: Notation

B Omitted proofs

B.1 Linearized version of Bernstein Condition

It will be convenient to work with the following *linearized version* of the Bernstein condition. It is an extension of a well-known result that has appeared in previous work, e.g. in [Koolen et al., 2016]. We restate it here for convenience.

Proposition 7 (Restatement of Proposition 5)**.** *Suppose that $(\mathcal{D}, \ell, \mathcal{F})$ satisfies the (B, β^*)-Bernstein condition for $\beta^* \in [0,1]$. Pick any $c > 0, \eta < 1/(2Bc)$. Then for all $0 < \beta \leq \beta^*$ and for all $f \in \mathcal{F}$:*

$$c \cdot \eta \mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}\left[(\ell(f; Z') - \ell(f^*; Z'))^2\right] \leq \left(\frac{1}{2} \wedge \beta\right) \cdot \left(\mathop{\mathbb{E}}_{Z' \sim \mathcal{D}}[\ell(f; Z') - \ell(f^*; Z')]\right) + (1 - \beta) \cdot (2Bc\eta)^{\frac{1}{1-\beta}}$$

Proof of Proposition 5. We first prove the proposition for $0 < \beta < 1$. For any $\eta > 0$, $B' > 0$, let $g(x) = B'\eta x^\beta - x$, for $x > 0$. We have

$$\max_{x>0}\{g(x)\} = \max_{x>0}\{B'\eta x^\beta - x\} = (1 - \beta)\beta^{\beta/(1-\beta)} \cdot (B'\eta)^{1/(1-\beta)},$$

since $g'(x) = 0$ for $x = (B'\eta\beta)^{1/(1-\beta)}$ and $g''(x) < 0$ for all $x > 0$. Hence, for all $0 < a \leq 1$ and $c > 0$, by setting $B' = Bc/a$, we have:

$$\begin{aligned}
\max_{x>0}\{Bc\eta x^{\beta} - ax\} &= \max_{x>0}\{a \cdot g(x)\} \\
&= a \cdot (1-\beta) \cdot \beta^{\beta/(1-\beta)} \cdot \left(\frac{Bc}{a}\eta\beta\right)^{1/(1-\beta)} \\
&= a^{-\beta/(1-\beta)} \cdot (1-\beta) \cdot \beta^{\beta/(1-\beta)} \cdot (Bc\eta)^{1/(1-\beta)} \qquad (23)
\end{aligned}$$

Now, by assumption, the (B, β^*)-Bernstein condition holds for $\beta^* \geq \beta$. Since the excess risk $R(f; \mathcal{D}) = \mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}[\ell(f; Z') - \ell(f^*; Z')] \in [0, 1]$, the (B, β)-Bernstein condition also holds, which implies that

$$c\eta \mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}\Big[\left(\ell(f; Z') - \ell(f^*; Z')\right)^2\Big] \leq Bc\eta\Big(\mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}[\ell(f; Z') - \ell(f^*; Z')]\Big)^{\beta}.$$

We now apply (23) with $x = \mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}[\ell(f; Z') - \ell(f^*; Z')]$ and $a = \left(\frac{1}{2} \wedge \beta\right)$ in the above inequality, establishing that

$$c\eta \mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}\Big[\left(\ell(f; Z') - \ell(f^*; Z')\right)^2\Big] \leq a \mathop{\mathbb{E}}_{Z'\sim\mathcal{D}}[\ell(f; Z') - \ell(f^*; Z')] + a^{-\frac{\beta}{1-\beta}} \cdot (1-\beta) \cdot \beta^{\frac{\beta}{1-\beta}} \cdot (Bc\eta)^{\frac{1}{1-\beta}}.$$

Bounding the last term of the RHS by $(1-\beta) \cdot (2Bc\eta)^{1/(1-\beta)}$ would complete the proof for $0 < \beta < 1$. For this to hold, it suffices to prove that $(\beta/a)^{\beta/(1-\beta)} \leq 2^{1/(1-\beta)}$, for $0 < \beta < 1$. If $a = \beta$, then the inequality reduces to $1 \leq 2$, which trivially holds. If $a = 1/2$, then the inequality reduces to $(2\beta)^{\beta} \leq 2$, which also holds.

It remains to prove the proposition for the limiting cases of $\beta = 0$ and $\beta = 1$. For $\beta = 0$, the RHS reduces to $(2Bc\eta)$, and the inequality trivially holds by the assumption of the (B, β^*)-Bernstein condition and the trivial bound of $(\mathbb{E}_{Z'\sim\mathcal{D}}[\ell(f; Z') - \ell(f^*; Z')])^{\beta^*} \leq 1$. For $\beta = \beta^* = 1$, the RHS reduces to $\frac{1}{2}\mathbb{E}_{Z'\sim\mathcal{D}}[\ell(f; Z') - \ell(f^*; Z')]$, and the inequality also holds by the assumption of the $(B, 1)$-Bernstein condition and our setting of $\eta < 1/(2Bc)$. □

B.2 Proof of main technical Lemma 1

For convenience we first restate the lemma:

Lemma 2 (Restatement of main technical Lemma 1)**.** *Fix any two real numbers r_0, r_1 such that $|r_0|, |r_1| \leq 1$. Let $S \sim \mathrm{Ber}(1/2)$ and let $\bar{S} = 1 - S$. Then for all $0 < \eta < 1/(1+\sqrt{2})$, it holds that*

$$r_{\bar{S}} - r_S \trianglelefteq_{\eta} \eta \cdot C_{2,\eta} r_{\bar{S}}^2$$

with $C_{A,\eta}$ an increasing function of η given by

$$C_{A,\eta} = \frac{1}{1-\eta} \cdot \left(A + \sqrt{A} \cdot \frac{\eta}{1-\eta} \cdot c_{\sqrt{A}\eta/(1-\eta)}\right),$$

where $c_{\gamma} = 2\frac{(-\log(1-\gamma)-\gamma)}{\gamma^2}$. If both r_0 and r_1 have the same sign, the constant can be improved to $C_{1,\eta}$ and the result holds for all $0 < \eta < 1/2$. Since c_{γ} is increasing and $\lim_{\gamma\downarrow 0} c_{\gamma} = 1$ the 'leading constant' is given by $\lim_{\eta\downarrow 0} C_{2,\eta} = 2$ (and $\lim_{\eta\downarrow 0} C_{1,\eta} = 1$ in case both r_0 and r_1 have the same sign).

For simplicity in the derivations, in the main text we will consider only $\eta \leq 1/4$ and use $C_{1/4} = 3.6064$ as an upper bound on $C_{2,\eta}$. It is easy to see that the result is tight in the limit for $\eta \downarrow 0$, by considering the case $r_0 = -r_1$ and doing a second order Taylor approximation of $\mathbb{E}_S[\exp(\eta(r_{\bar{S}} - r_S))]$ around $\eta = 0$. The result is only proven for r_0, r_1 with $|r_0|, |r_1| \leq 1$, and (since c_{γ} tends to ∞ as $\gamma \uparrow 1$), the bound becomes void for $\eta \geq 1/(1+\sqrt{2})$. Yet, as is straightforward to show by inspecting the formulas, for general $r_0 \leq r_1 < \infty$ we still have $r_{\bar{S}} - r_S \trianglelefteq_{\eta} \eta B r_{\bar{S}}^2$ for some finite B as long as $r_0 > -\log 2$, with B tending to infinity as $r_0 \downarrow -\log 2$; it is just not so easy any more to give a crisp bound.

The proof crucially makes use of the following *un-expected Bernstein inequality* (originally due to Fan et al. [2015], our presentation follows Mhammedi et al. [2019] who gave it its name):

Lemma 3 (Un-expected Bernstein Inequality [Mhammedi et al., 2019, Lemma 13(a)])**.** *Let U be a random variable bounded from above by $b > 0$ almost surely, and let $\theta(u) = (-\log(1-u) - u)/u^2$. For all $0 < \eta < 1/b$, we have*

$$\mathbb{E}[U] - U \trianglelefteq_{\eta} \frac{1}{2}\eta c_{\eta} \cdot U^2 \quad \textit{for all } c_{\eta} \geq 2 \cdot \theta(\eta b).$$

Proof of Lemma 1. We only prove the case for general r_0, r_1 with $|r_0|, |r_1| \leq 1$. The improved bound for r_0 and r_1 of the same sign can be proven by following exactly the same steps as below, where the term $(r_0 - r_1)^2$ in the derivation of (26) is bounded by $r_1^2 + r_2^2$ instead of $2r_1^2 + 2r_0^2$.

Fix $\lambda > 0$ and let $x \in \mathbb{R}$. The well-known cosh-inequality states that $(1/2)\exp(\lambda x) + (1/2)\exp(-\lambda x) \leq \exp(\lambda^2 x^2/2)$. Now fix x and let Y be a Rademacher RV such that $P(Y = x) = P(Y = -x) = 1/2$. By definition, for all $\lambda > 0$, $\mathbb{E}_{Y}[\exp(\lambda Y)] = (1/2)\exp(\lambda x) + (1/2)\exp(-\lambda x)$. Therefore, by the cosh-inequality, we have for all $\eta > 0, A > 0$, and letting $\lambda = A\eta$, that

$$Y \trianglelefteq_{A\eta} \frac{1}{2} A\eta x^2. \tag{24}$$

Now, let U be a RV such that $U \in [0, 1]$. Then by the *un-expected Bernstein inequality* of Mhammedi et al. [2019] (Lemma 3) we have, for all $0 < \eta < 1$,

$$\mathop{\mathbb{E}}_{U}[U] \trianglelefteq_{\eta} U + \frac{1}{2}\eta c_{\eta} U^2,$$

for $c_{\eta} = 2\frac{(-\ln(1-\eta)-\eta)}{\eta^2}$. Since $U \geq 0$, it follows that for all $0 < \eta < 1$,

$$\mathop{\mathbb{E}}_{U}[U] \trianglelefteq_{\eta} (1 + \eta c_{\eta}/2) U.$$

Hence

$$2A\eta \mathop{\mathbb{E}}_{U}[U] \trianglelefteq_{1/2A} A\eta(2 + \eta c_{\eta}) U. \tag{25}$$

Note that with $x = r_1 - r_0$, $r_{\bar{S}} - r_S$ is a Rademacher RV such that $P(r_{\bar{S}} - r_S = x) = P(r_{\bar{S}} - r_S = -x) = \frac{1}{2}$. Thus, by (24), we have that for all $\eta > 0$, $A > 0$,

$$\begin{aligned} r_{\bar{S}} - r_S &\trianglelefteq_{A\eta} \frac{1}{2} A\eta \cdot (r_0 - r_1)^2 \\ &\leq A\eta \cdot \frac{1}{2}(2r_1^2 + 2r_0^2) && \text{(since } (x-y)^2 \leq 2x^2 + 2y^2\text{)} \\ &= 2A\eta \cdot \left(\mathop{\mathbb{E}}_{S'}\left[r_{\bar{S}'}^2 \right] \right). && (26) \end{aligned}$$

Since $r_{\bar{S}}^2 \in [0, 1]$, we also apply (25) to $r_{\bar{S}}^2$. We then have that, for $\eta < 1$,

$$2A\eta \cdot \left(\mathop{\mathbb{E}}_{S'}\left[r_{\bar{S}'}^2 \right] \right) \trianglelefteq_{A\eta}^{S'} A\eta(2 + \eta c_{\eta}) r_{\bar{S}}^2.$$

Now for arbitrary (possibly dependent) RVs X, Y, Z we have $X \trianglelefteq_{A\eta} Y, Y \trianglelefteq_{1/2A} Z \Rightarrow X \trianglelefteq_{\bar{\eta}} Z$, where $\bar{\eta} = (1/(A\eta) + 2A)^{-1} = A\eta/(1 + 2A^2\eta)$ (by Proposition 2). Combining the above two ESIs implies that

$$r_{\bar{S}} - r_S \trianglelefteq_{\bar{\eta}} A\eta(2 + \eta c_{\eta}) r_{\bar{S}}^2.$$

This bound holds for all $0 < \eta < 1$ and arbitrary $A > 0$. We want this bound to hold for as large $\bar{\eta}$ as possible. Since $\eta = \bar{\eta}/(A(1 - 2A\bar{\eta}))$ is an increasing function of $\bar{\eta}$, the bound is valid up to all $\bar{\eta} < \bar{\eta}^*$ where $1 = \bar{\eta}^*/(A(1 - 2A\bar{\eta}^*))$. Choosing the A for which $\bar{\eta}^*$ is maximal gives $A = 1/\sqrt{2}$, and then $\bar{\eta}^* = 1/(1 + \sqrt{2})$ and $\eta = \sqrt{2}\bar{\eta}/(1 - \bar{\eta})$. Substituting η and A in the previous ESI we now get, for $0 < \bar{\eta} < 1/(1 + \sqrt{2})$,

$$r_{\bar{S}} - r_S \trianglelefteq_{\bar{\eta}} \frac{\bar{\eta}}{1 - \bar{\eta}} \cdot \left(2 + \sqrt{2}\frac{\bar{\eta}}{1 - \bar{\eta}} \cdot c_{\sqrt{2}\bar{\eta}/(1-\bar{\eta})} \right) \cdot r_{\bar{S}}^2 = \bar{\eta} \cdot C_{\sqrt{2}, \bar{\eta}} \cdot r_{\bar{S}}^2$$

and the result follows. □

B.3 Improved in-expectation bound - 'Variation' of Theorem 1

Corollary 4. ('**Variation** of Theorem 1' - Restatement of Corollary 2) *Consider the setting and notation of Theorem 1. For all $\beta \in [0, \beta^*]$, it holds that*

$$\mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}}\left[L(A|\tilde{Z}_\mathbf{0}; \mathcal{D}) - L(A|\tilde{Z}_\mathbf{0}; \tilde{Z}_\mathbf{0})\right] \leq$$

$$(1 \wedge 2\beta) \cdot \mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}}\left[R(A|\tilde{Z}_\mathbf{0}; \tilde{Z}_\mathbf{0})\right] + 4 \cdot \left(\frac{\mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}, \tilde{Z}_\mathbf{1}}\left[\mathsf{KL}\left(A|\tilde{Z}_\mathbf{0} \middle\| \pi|\langle \tilde{Z}_\mathbf{0}, \tilde{Z}_\mathbf{1}\rangle\right)\right]}{n\eta_{\max}}\right)_{[**]}^{\frac{1}{2-\beta}}. \tag{27}$$

The proof follows by a few modifications of the proof of the main Theorem 1.

Proof Sketch. The proof would be the same up to and including the derivation of inequality (14), where $\eta < \eta_{\max}$ is not random. At this step, we can weaken this ESI to an in-expectation inequality, subsequently derive and add the equivalent of inequalities (18) and (19), to yield

$$\mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}}\left[L(A|\tilde{Z}_\mathbf{0}; \mathcal{D}) - L(A|\tilde{Z}_\mathbf{0}; \tilde{Z}_\mathbf{0})\right] \leq$$

$$(1 \wedge 2\beta) \cdot \mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}}\left[R(A|\tilde{Z}_\mathbf{0}; \tilde{Z}_\mathbf{0})\right] + 2 \cdot \left(\left(\frac{\eta}{\eta_{\max}}\right)^{\frac{1}{1-\beta}} + \frac{\mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}, \tilde{Z}_\mathbf{1}}\left[\mathsf{KL}\left(A|\tilde{Z}_\mathbf{0} \middle\| \pi|\langle \tilde{Z}\rangle\right)\right]}{n\eta}\right).$$

By differentiation, we choose $\eta = \left(\eta_{\max} \wedge (1-\beta)^{\frac{1-\beta}{2-\beta}} \eta_{\max}^{\frac{1}{2-\beta}} \left(\frac{\mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}, \tilde{Z}_\mathbf{1}}\left[\mathsf{KL}\left(A|\tilde{Z}_\mathbf{0} \middle\| \pi|\tilde{Z}\right)\right]}{n\eta_{\max}}\right)^{\frac{1-\beta}{2-\beta}}\right)$ to minimize the sum of the last two terms of the RHS of the inequality, which gives the improved in-expectation bound:

$$\mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}}\left[L(A|\tilde{Z}_\mathbf{0}; \mathcal{D}) - L(A|\tilde{Z}_\mathbf{0}; \tilde{Z}_\mathbf{0})\right] \leq (1 \wedge 2\beta) \cdot \mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}}\left[R(A|\tilde{Z}_\mathbf{0}; \tilde{Z}_\mathbf{0})\right] + 4 \cdot \left(\frac{\mathop{\mathbb{E}}_{\tilde{Z}_\mathbf{0}, \tilde{Z}_\mathbf{1}}\left[\mathsf{KL}\left(A|\tilde{Z}_\mathbf{0} \middle\| \pi|\tilde{Z}\right)\right]}{n\eta_{\max}}\right)_{[**]}^{\frac{1}{2-\beta}},$$

where $a^b_{[**]} = \max\{a^b, a\}$. □

B.4 Proof of Theorem 2 (VC classes)

First, we formally define the global consistency property. Here we abuse notation by interchanging between $(\mathcal{X} \times \mathcal{Y})^n$ and $\mathcal{X}^n \times \mathcal{Y}^n$. That is, we refer to $(x, y) \in (\mathcal{X} \times \mathcal{Y})^n$ when we mean $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}^n$. We also use (and abuse) the notation $(\mathcal{X} \times \mathcal{Y})^* := \bigcup_{n=0}^{\infty} (\mathcal{X} \times \mathcal{Y})^n$. Thus the notation $(x, y) \in (\mathcal{X} \times \mathcal{Y})^*$ means, for some n, we have $x \in \mathcal{X}^n$ and $y \in \mathcal{Y}^n$.

Definition 10 (Global Consistency Property)**.** *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathcal{Y}$. A deterministic algorithm $A : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{F}$ is said to have the global consistency property if the following holds. Let $(x, y) \in (\mathcal{X} \times \mathcal{Y})^*$ and let $f = A|(x, y)$. We require that, for any $x' \in \mathcal{X}^*$ such that x' contains all the elements of x (i.e., $\forall i \; \exists j \; x_i = x'_j$), we have $A|(x', y') = f$ whenever $y'_i = f(x'_i)$ for all i.*

Informally, this property says the following. Suppose the algorithm is run on some labelled dataset (x, y) to obtain an output hypothesis $f = A|(x, y)$. If the dataset is relabelled to be perfectly consistent with f, then the algorithm should still output f. This should also hold if further examples are added to the dataset (where the additional examples are also consistent with f).

The proof of the theorem is split in the next two lemmata.

Lemma 4. *Let $A : (\mathcal{X} \times \{0,1\})^n \to \mathcal{F}$ be a deterministic algorithm, where $\mathcal{F}$ is a class of functions $f : \mathcal{X} \to \{0,1\}$ with VC dimension d. Suppose A (appropriately extended to inputs of arbitrary size) has the global consistency property. Then for any $\tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}} \in \mathcal{Z}^n$,*

$$\mathsf{KL}(A|\tilde{z}_{\mathbf{0}} \| \pi|\langle \tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}}\rangle) \leq d \log(2n).$$

Lemma 5. *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \{0,1\}$. Then there exists a deterministic algorithm $A : (\mathcal{X} \times \{0,1\})^* \to \mathcal{F}$ that has the global consistency property and is an empirical risk minimizer – that is, for all $(x,y) \in (\mathcal{X} \times \{0,1\})^*$, if $f^* = A|(x,y)$, then*

$$\sum_i \mathbb{I}[f^*(x_i) \neq y_i] = \min_{f \in \mathcal{F}} \sum_i \mathbb{I}[f(x_i) \neq y_i].$$

To prove Lemma 4 we invoke the Sauer-Shelah lemma:[3]

Lemma 6 (Sauer [1972], Shelah [1972])**.** *Let $\mathcal{F}$ be a class of functions $f : \mathcal{F} \to \{0,1\}$ with VC dimension d. For any $x = \{x_1, \cdots, x_m\} \subset \mathcal{X}$, the number of possible labellings of x induced by $\mathcal{F}$ is*

$$|\{(f(x_1), f(x_2), \cdots, f(x_m)) : f \in \mathcal{F}\}| \leq \sum_{k=0}^{d} \binom{m}{k} \leq \begin{cases} (em/d)^d & \text{if } m \geq d \\ e^2 \cdot (m/2)^d & \text{if } m \geq 2 \\ e \cdot m^d & \text{if } m \geq 1 \end{cases}.$$

Here we define $\binom{m}{k} = 0$ if $k > m$. Thus $\sum_{k=0}^{d} \binom{m}{k} = 2^m$ if $m \leq d$. Note that we give three different forms of the final bound for convenience, all of which are derived from the bound

$$\forall m \geq d \quad \forall x \geq 1 \qquad \sum_{k=0}^{d} \binom{m}{k} \leq \sum_{k=0}^{d} \binom{m}{k} x^{d-k} \leq \sum_{k=0}^{m} \binom{m}{k} x^{d-k} = \left(1 + x^{-1}\right)^m \cdot x^d \leq e^{m/x} \cdot x^d.$$

Proof of Lemma 4. Let $\tilde{z} = (\tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}})$ be the fixed supersample and let $\tilde{x} = \{x : \exists y \in \{0,1\}, i \in [n], j \in \{0,1\} : (x,y) = \tilde{z}_{i,j}\}$ be the set of all unlabelled examples in $\tilde{z}$. We choose as an almost exchangeable prior distribution π the following: $\pi(f) = \mathbb{I}\{\exists s \in \{0,1\}^n : A|\tilde{z}_s = f\}/|H(\tilde{z})|$, where $H(\tilde{z}) = \{A|\tilde{z}_s \text{ for some } s \in \{0,1\}^n\}$. That is, π is uniform over all the possible outputs of algorithm A given input $\tilde{z}_s$ for some $s \in \{0,1\}^n$. Then the KL term is written as

$$\mathsf{KL}(A|\tilde{z}_{\mathbf{0}} \| \pi|\langle \tilde{z}_{\mathbf{0}}, \tilde{z}_{\mathbf{1}}\rangle) = \log \frac{1}{\pi(A|\tilde{z}_{\mathbf{0}})} = \log \frac{|H(\tilde{z})|}{\mathbb{I}\{\exists s \in \{0,1\}^n : A|\tilde{z}_s = A|\tilde{z}_{\mathbf{0}}\}} = \log |H(\tilde{z})|.$$

It suffices to bound $|H(\tilde{z})|$. By the global consistency property, if $A|\tilde{z}_s = f$ for some $s \in \{0,1\}^n$, then it must be that $A|(\tilde{x}, f(\tilde{x})) = f$. Therefore

$$H(\tilde{z}) \subseteq \{A|(\tilde{x}, f(\tilde{x})) : f \in \mathcal{F}\} \subseteq \{f(\tilde{x}) : f \in \mathcal{F}\}$$

By Lemma 6, the set of all the possible labellings of $\tilde{x} \in \mathcal{X}^{2n}$ by $\mathcal{F}$ has size at most $|\{f(\tilde{x}) : f \in \mathcal{F}\}| \leq (2n)^d$. Thus, $|H(\tilde{z})| \leq (2n)^d$ and the bound of the lemma follows. □

Lemma 5 is exaclty the same as the corresponding lemma in the proof of the CMI result of Steinke and Zakynthinou [2020]. We present their proof here to give a clear picture of the type of algorithm that could satisfy the lemma for our examples.

For the proof, we will invoke the well-ordering theorem Zermelo [1904]:

Lemma 7 (Zermelo [1904])**.** *Let $\mathcal{F}$ be a set. Then there exists a binary relation $\preceq$ with the following properties.*

- *Transitivity:* $\forall f, g, h \in \mathcal{F} \quad f \preceq g \wedge g \preceq h \implies f \preceq h$
- *Totality:* $\forall f, g \in \mathcal{F} \quad f \preceq g \vee g \preceq f$

[3] Vapnik and Chervonenkis proved a slightly weaker bound, namely $|\{(f(x_1), f(x_2), \cdots, f(x_m)) : f \in \mathcal{F}\}| \leq m^{d+1} + 1$ for $m > d$ [Vapnik and Chervonenkis, 1971, Thm. 1].

- *Antisymmetry:* $\forall f, g \in \mathcal{F} \quad f \preceq g \wedge g \preceq f \Leftrightarrow f = g$
- *Well-order:* $\forall H \subset \mathcal{F} \;\; (\; H \neq \emptyset \quad \implies \quad \exists h \in H \;\; \forall f \in H \;\; h \preceq f \;)$

Let $\preceq$ be a well-ordering of $\mathcal{F}$. On a finite computer, we could simply let $\preceq$ be the lexicographical ordering on the binary representations of elements of $\mathcal{F}$.

Proof of Lemma 5. An empirical risk minimizer $A : (\mathcal{X} \times \{0,1\})^n \to \mathcal{F}$ must have the property

$$\forall (x,y) \in (\mathcal{X} \times \{0,1\})^n \qquad A|(x,y) \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \ell(f,(x,y)) := \left\{ f \in \mathcal{F} : \ell(f,(x,y)) = \inf_{f' \in \mathcal{F}} \ell(f',(x,y)) \right\}.$$

However, we must also ensure that A satisfies the global consistency property. The only difficulty that arises here is when the argmin contains multiple hypotheses; we must break ties in a consistent manner. (Note that the argmin is never empty, as the 0-1 loss $\ell(f',(x,y)) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}[f'(x_i) \neq y_i]$ always takes values in the finite set $\{0, 1/n, 2/n, 3/n, \cdots, 1\}$.)

Whenever there are multiple $f \in \mathcal{F}$ that minimize $\ell(f,(x,y))$, our algorithm $A|(x,y)$ chooses the least element according to the well-ordering. In symbols, A satisfies the following two properties, which also uniquely define it.

$$\forall (x,y) \in (\mathcal{X} \times \{0,1\})^* \;\; \forall h \in \mathcal{F} \quad \left(\begin{array}{c} \ell(A|(x,y),(x,y)) \leq \ell(f,(x,y)) \\ \wedge \\ \ell(A|(x,y),(x,y)) = \ell(f,(x,y)) \implies A|(x,y) \preceq f \end{array} \right).$$

By construction, our algorithm A is an empirical risk minimizer. It only remains to prove that it satisfies the global consistency property. To this end, let $(x,y) \in (\mathcal{F} \times \{0,1\})^n$ and let $x' \in \mathcal{X}^m$ where x' contains all the elements of x (i.e., $\forall i \in [n] \; \exists j \in [m] \; x_i = x'_j$). Let $f = A|(x,y)$ and $f' = A|(x', f(x'))$. We must prove that $f' = f$.

By construction, the empirical loss of f on the dataset $(x', f(x'))$ is 0. Since f' is the output of an empirical risk minimizer on the dataset $(x', f(x'))$, it too has empirical loss 0 on this dataset. In particular, $f(x'_j) = f'(x'_j)$ for all $j \in [m]$. Moreover, since A breaks ties using the ordering, we have $f' \preceq f$. However, since f and f' agree on x', they also agree on x and, hence, have the same loss on the dataset (x,y) – that is, $\ell(f',(x,y)) = \ell(f,(x,y)) = \inf_{hf' \in \mathcal{F}} \ell(f'',(x,y))$. This means that $A|(x,y)$ outputting f implies that $f \preceq f'$. Thus we conclude that $f = f'$, as required. □

B.5 Proof Sketch for Gibbs example

Proof Sketch. A known property of the $\hat{\eta}$-Gibbs algorithm (see for example [Grünwald and Mehta, 2020]) relative to prior $W \mid \tilde{z}_{\mathbf{0}}$ is that, among all learning algorithms A that output a distribution on $\mathcal{F}$, for all $\tilde{z}_{\mathbf{0}}$ it achieves

$$\min_A R(A|\tilde{z}_{\mathbf{0}}; \tilde{z}_{\mathbf{0}}) + \frac{\mathsf{KL}(A|\tilde{z}_{\mathbf{0}} \| W|\tilde{z}_{\mathbf{0}})}{n\hat{\eta}}. \tag{28}$$

Now assume that the prior W is a compression scheme prior of some size k and let $\pi|\langle\cdot\rangle$ denote the corresponding almost exchangeable prior satisfying (22). If we consider the proof of Theorem 1 again, we see that if we set $A := A_{\text{GIBBS}}$ to the Gibbs algorithm, and A' to any other learning algorithm, then the crucial inequality (15) in the proof of Theorem 1 still holds with $R(A|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}})$ on the right-hand side replaced by $R(A'|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}})$ and UB set to $\mathsf{KL}\left(A'|\tilde{Z}_{\mathbf{0}} \middle\| W|\tilde{Z}_{\mathbf{0}}\right) + k \log 2n$. Following all the remaining steps in the proof while keeping UB in its new definition and keeping the distinction between A' and A, we get the following corollary of Theorem 1: *if we set A to the Gibbs algorithm relative to size k compression scheme prior W, and A' to any other algorithm, we have, with the same abbreviations as in Theorem 1*

$$L(A_{\text{GIBBS}}|\tilde{Z}_{\mathbf{0}}; \mathcal{D}) \trianglelefteq_{\eta}^{\tilde{Z}_{\mathbf{0}}}$$
$$L(A'|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}}) + (1 \wedge 2\beta) \cdot R(A'|\tilde{Z}_{\mathbf{0}}; \tilde{Z}_{\mathbf{0}}) + 8 \cdot \left(\frac{\mathsf{KL}\left(A'|\tilde{Z}_{\mathbf{0}} \middle\| W|\tilde{Z}_{\mathbf{0}}\right) + O(k \log n)}{n\eta_{\max}} \right)^{1/(2-\beta)}_{[**]} + \frac{6\eta}{n}.$$

□