PAC-BAYES AND INFORMATION COMPLEXITY

Pradeep Kr. Banerjee
MPI MiS
pradeep@mis.mpg.de

Guido Montúfar
UCLA & MPI MiS
montufar@math.ucla.edu

ABSTRACT

We point out that a number of well-known PAC-Bayesian-style and information-theoretic generalization bounds for randomized learning algorithms can be derived under a common framework starting from a fundamental information exponential inequality. We also obtain new bounds for data-dependent priors and unbounded loss functions. Optimizing these bounds naturally gives rise to a method called Information Complexity Minimization for which we discuss two practical examples for learning with neural networks, namely Entropy- and PAC-Bayes- SGD.

1 INTRODUCTION

A fundamental observation in statistical learning theory is that information compression and learning are intrinsically related in the sense that both entail identifying statistical regularities and patterns in the data. This relation has been formalized over the years in various ways, e.g., via sample compression schemes (Littlestone & Warmuth, 1986; Moran & Yehudayoff, 2016), Occam's razor (Blumer et al., 1987) and minimum description length arguments (Li et al., 2003; Blum & Langford, 2003), and more recently via different notions of information stability (Russo & Zou, 2016; Xu & Raginsky, 2017; Jiao et al., 2017; Bassily et al., 2018; Bu et al., 2019; Steinke & Zakynthinou, 2020a;b). Information stability quantifies the sensitivity of a learning algorithm to local perturbations of its input, i.e., the training data, and draws on a rich tradition of earlier work on algorithmic (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010; Bassily et al., 2016) and distributional (Dwork et al., 2015; Rogers et al., 2016; Feldman & Steinke, 2018) stability in adaptive data analysis.

On the other hand, an important approach to data-dependent generalization bounds is PAC-Bayes, originally due to McAllester (1999a;b; 2013). While PAC-Bayes and information stability have evolved independently of each other, a principal objective of this note is to present them under a unified framework. We show that the information exponential inequality (IEI) due to Zhang (2006a) gives a general recipe for constructing bounds of both flavors. Besides recovering several important bounds such as the mutual information-bound due to Xu & Raginsky (2017) and the classical PAC-Bayesian bounds due to Catoni (2007) and McAllester (2013), we also obtain new bounds for data-dependent priors and unbounded loss functions. Optimizing these bounds gives rise to variants of the Gibbs algorithm, for which we discuss two examples for learning with neural networks, namely, Entropy-SGD (Chaudhari et al., 2017) and PAC-Bayes-SGD (Dziugaite & Roy, 2017). We also show a PAC-Bayes bound motivated by an Occam's factor argument, which can be interpreted in relation to flat minima (Hochreiter & Schmidhuber, 1997).

Three key ideas guide our discussion, namely, (1) the lesser the information revealed by an algorithm about its input, the better the generalization, (2) data-dependent priors entail tighter generalization bounds, and (3) optimizing such bounds is a natural recipe for designing new learning algorithms.

2 THE INFORMATION EXPONENTIAL INEQUALITY AND APPLICATIONS

We consider the standard apparatus of statistical learning theory (Shalev-Shwartz & Ben-David, 2014). We have an example domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ of the instances and labels, a hypothesis space $\mathcal{W}$, a fixed loss function $\ell : \mathcal{W} \times \mathcal{Z} \to [0, \infty)$, and a training sample S, which is an n-tuple $(Z_1, \dots, Z_n)$ of i.i.d. random elements of $\mathcal{Z}$ drawn according to some unknown distribution μ. A learning algorithm is a Markov kernel $P_{W|S}$ that maps input training samples S to conditional distributions of hypotheses W in $\mathcal{W}$. For given n, this defines a joint distribution $P_{SW} = P_S P_{W|S}$, $P_S = \mu^{\otimes n}$, and a corresponding marginal distribution P_W. The *true risk* of a hypothesis $w \in \mathcal{W}$ on μ is $L_\mu(w) := \mathbb{E}_\mu[\ell(w, Z)]$, and its *empirical risk* on the training sample S is $L_S(w) := \frac{1}{n} \sum_{i=1}^n \ell(w, Z_i)$. Our goal is to control the *generalization error*, $\mathrm{g}(W, S) := L_\mu(W) - L_S(W)$, either in expectation or with high probability.

For controlling the generalization error in expectation, we can rewrite the true risk of a given hypothesis w as $L_\mu(w) = \mathbb{E}_{S' \sim \mu^{\otimes n}}[L_{S'}(w)]$, where $S' = (Z'_1, \ldots, Z'_n)$ is an i.i.d. sample. Then the expected generalization error can be written as a difference of two expectations of the same loss function, $\mathbb{E}_{SW}[\mathrm{g}(W,S)] = \mathbb{E}_{P_S \otimes P_W}[L_S(W)] - \mathbb{E}_{P_{SW}}[L_S(W)]$, where the second expectation is taken w.r.t. the joint distribution of the training sample and the output hypothesis, while the first expectation is taken w.r.t. the product of the two marginal distributions. Hence the expected generalization error reflects the dependence of the output W on the input S. This dependence can also be measured by their mutual information as has been shown in recent works (Xu & Raginsky, 2017; Jiao et al., 2017; Bassily et al., 2018). We refer to such bounds as mutual information-based generalization bounds.

Alternatively, we may wish to control the generalization error of the learning algorithm $P_{W|S}$ with high probability over the training sample S. The expected generalization error over hypotheses chosen from the distribution P (*posterior*) output by the learning algorithm, i.e., $\mathbb{E}_P[\mathrm{g}(W,S)]$, can be upper-bounded with high probability under P_S by the KL divergence between P and an arbitrary reference distribution Q (*prior*), that is selected *before* the draw of the training sample S. For any Q, these bounds hold uniformly for all P, and are called PAC-Bayesian bounds (McAllester, 1999a;b; 2013; Maurer, 2004; Zhang, 2006a; Catoni, 2007; Alquier et al., 2016; Germain et al., 2009; 2016; London, 2017; Thiemann et al., 2017; Grünwald & Mehta, 2020; Rivasplata et al., 2020), where PAC stands for *Probably Approximately Correct*. Bounds of this type are useful when we have a fixed dataset $s \in \mathcal{Z}^n$ and a new hypothesis is sampled from P every time the algorithm is used. Choosing the posterior to minimize a PAC-Bayesian bound leads to the well-known Gibbs algorithm (Zhang, 2006a; Xu & Raginsky, 2017; Alquier et al., 2016; Kuzborskij et al., 2019). On the other hand, for a fixed posterior P, $\mathbb{E}_S[D(P\|Q)]$ is minimized by the *oracle prior*, $Q^\star = \mathbb{E}_S[P_{W|S}(\cdot|S)]$. Note $\mathbb{E}_S[D(P\|Q^\star)]$ is just the mutual information $I(S;W)$, which is the key quantity controlling the expected generalization error in (Xu & Raginsky, 2017; Jiao et al., 2017; Bassily et al., 2018).

For any $\beta > 0$, we define the *annealed expectation*, $M_\beta(w) = -\beta^{-1} \ln \mathbb{E}_\mu[e^{-\beta \ell(w,Z)}]$, which acts as a surrogate for $L_\mu(w)$. We write $\mathcal{M}(\mathcal{W})$ to denote the family of probability measures over a set $\mathcal{W}$. The following fundamental inequality is due to Tong Zhang:

Lemma 1 (Information exponential inequality (IEI), Zhang 2006a)**.** *For any prior $Q \in \mathcal{M}(\mathcal{W})$, any real-valued loss function ℓ on $\mathcal{W} \times \mathcal{Z}$, and any posterior distribution $P \ll Q$ over $\mathcal{W}$ that depends on an i.i.d. training sample S, we have $\mathbb{E}_S \exp\left\{n\beta \mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)\right\} \leq 1$.*

The IEI implies bounds both in probability and in expectation for the quantity $n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)$, and is the key tool for showing the following result:

Theorem 2. *Let μ be a distribution over $\mathcal{Z}$, and let S be an i.i.d. training sample from μ. Let $Q \in \mathcal{M}(\mathcal{W})$ be a prior distribution that does not depend on S, and let ℓ be a real-valued loss function on $\mathcal{W} \times \mathcal{Z}$. Suppose that there exist a convex function $\psi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfying $\psi(0) = \psi'(0) = 0$, such that $\sup_{w \in \mathcal{W}}[L_\mu(w) - M_\beta(w)] \leq \frac{\psi(\beta)}{\beta}$, $\forall \beta > 0$. Then, for any $\beta > 0$, and $\delta \in (0,1]$, with probability of at least $1-\delta$ over the choice of $S \sim \mu^{\otimes n}$, for all distributions $P \ll Q$ over $\mathcal{W}$ (even such that depend on S), we have*

$$\mathbb{E}_P[\mathrm{g}(W,S)] \leq \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right) + \frac{\psi(\beta)}{\beta}. \tag{1}$$

Moreover, we have the following bound in expectation:

$$\mathbb{E}_{SW}[\mathrm{g}(W,S)] \leq \psi^{*-1}\left(\frac{D(P\|Q|P_S)}{n}\right), \tag{2}$$

where ψ^{*-1} is the inverse of the Fenchel-Legendre dual of ψ.

The proof of Theorem 2 is given in Appendix A.2. Under the oracle prior $Q^\star = \mathbb{E}_S[P_{W|S}]$, $\mathbb{E}_S[D(P\|Q^\star)] = I(S;W)$. If $\ell(w,Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then we can take $\psi(\beta) = \beta^2\sigma^2/2$ and $\psi^{*-1}(y) = \sqrt{2\sigma^2 y}$ (Boucheron et al., 2013, §2.3), in which case we recover the bound in expectation due to Xu & Raginsky (2017):

Corollary 3. *If $\ell(w,Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then $\mathbb{E}_{SW}[\mathrm{g}(W,S)] \leq \sqrt{2\sigma^2 I(S;W)/n}$.*

Corollary 3 shows that *an algorithm that reveals a small amount of information about its input generalizes well*. This observation, for instance, forms the basis for the Gibbs algorithm, which can

be thought of as "stabilizing" the empirical risk minimization (ERM) algorithm by controlling the input-output mutual information $I(S;W)$ (Xu & Raginsky, 2017). We discuss extensions of this idea in Section 3. In Appendix A.3, we highlight a functional characterization of the mutual information in relation to the "single-draw" bound due to Bassily et al. (2018). In Appendix A.4, we show how several well-known PAC-Bayesian inequalities can be derived starting from the IEI. These include, for instance, the following classical bound due to Catoni:

Corollary 4 (Catoni 2007, Theorem 1.2.6)**.** *For any $\{0,1\}$-valued loss ℓ, any distribution μ, prior $Q \in \mathcal{M}(\mathcal{W})$, any real $\beta > 0$, and any $\delta \in (0,1]$, with probability of at least $1-\delta$ over $S \sim \mu^{\otimes n}$, we have for all $P \ll Q$ over $\mathcal{W}$:*

$$\mathbb{E}_P[L_\mu(W)] \leq \Phi_\beta^{-1}\Big\{\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\Big[D(P\|Q) + \ln\frac{1}{\delta}\Big]\Big\}, \text{ where } \Phi_\beta^{-1}(x) = \tfrac{1-e^{-\beta x}}{1-e^{-\beta}},\ x \in \mathbb{R}.$$

Other bounds include, for example, the "Linear PAC-Bayes bound" due to McAllester (2013, Theorem 2) and the "PAC-Bayes-KL inequality" (Seeger, 2002; Maurer, 2004) for $[0,1]$-valued loss functions, as well as the bounds for sub-Gaussian and sub-gamma loss functions due to Germain et al. (2016).

Differentially private data-dependent priors. A PAC-Bayesian bound such as (1) stipulates that the prior Q be chosen before the draw of the training sample S. Q may, however, depend on the data generating distribution μ (Lever et al., 2013). To have a good control over the KL term in (1), it is desirable that Q be "aligned" with the data-dependent posterior P. One way to achieve this goal is to choose Q based on S in a differentially private fashion so that Q is stable to local perturbations in S (Dziugaite & Roy, 2018b). We can then treat Q "as if" it is independent of S. The next result formalizes this argument and gives a PAC-Bayesian-style bound that is valid for data-dependent priors and unbounded loss functions.

Proposition 5. *Let $\mathcal{K}(\mathcal{S},\mathcal{W})$ denote the set of Markov kernels from $\mathcal{S}$ to $\mathcal{W}$. Let μ be a distribution over $\mathcal{Z}$, and let S be an i.i.d. training sample from μ. Let $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{W})$ be an $(\epsilon,0)$-differentially private algorithm and let ℓ be a real-valued loss function on $\mathcal{W}\times\mathcal{Z}$. Let $\beta > 0$, and let $\delta \in (0,1]$. Then with probability of at least $1-\delta$ over the choice of $S \sim \mu^{\otimes n}$, for all $P \in \mathcal{M}(\mathcal{W})$,*

$$\mathbb{E}_P[M_\beta(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q^0(S)) + \ln\frac{2}{\delta} + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\ln\frac{4}{\delta}}\right).$$

The proof of Proposition 5 is given in Appendix A.5. Proposition 5 is valid for any loss function and is similar in spirit to the traditional PAC-Bayesian bounds in (Dziugaite & Roy, 2018b, Theorem 4.2), and (Rivasplata et al., 2020, Eq. 7), which, however, either apply only when the loss is bounded in $[0,1]$, or entails approximating an exponential moment involving the true risk.

3 Information complexity minimization

Given a prior distribution, choosing a posterior to minimize a PAC-Bayesian bound naturally gives rise to a method called *Information Complexity Minimization (ICM)* (Zhang, 2006a;b). Concretely, for a given prior Q and hypothesis set $\mathcal{G} \subseteq \mathcal{M}(\mathcal{W})$, define the *Optimal Information Complexity (*OIC*)* at a given $\beta > 0$ as

$$\mathrm{OIC}_{\mathcal{G}}^{\beta} := \inf_{P\in\mathcal{G}}\Big\{\mathbb{E}_P[L_S(W)] + (n\beta)^{-1}D(P\|Q)\Big\}. \tag{3}$$

When $\mathcal{G} = \mathcal{M}(\mathcal{W})$, we recover the Gibbs algorithm, in which case the OIC evaluates to the *(extended) stochastic complexity*, $-\frac{1}{\beta}\ln\mathbb{E}_Q[e^{-\beta L_S(W)}]$ (Rissanen, 1989; Yamanishi, 1998). The latter in turn coincides with the log-Bayesian marginal likelihood for $\beta = 1$ and the log loss (Zhang, 2006a;b; Barron & Cover, 1991). We briefly discuss two examples of ICM for learning with neural networks.

PAC-Bayes-SGD. *PAC-Bayes-SGD* is an approach to computing nonvacuous generalization bounds for overparameterized neural network classifiers trained with stochastic gradient descent (SGD) (Dziugaite & Roy, 2017; Langford & Caruana, 2002; Hinton & van Camp, 1993). These bounds are obtained by *retraining* the network using an objective derived from a PAC-Bayes bound, starting from the solution found by SGD (or in fact any other procedure) for the training loss $L_S(w)$ w.r.t. w. We show how Catoni's bound in Corollary 4 can be used to derive a PAC-Bayes-SGD objective.

Consider a binary classification setting with examples domain $\mathcal{Z} = \mathcal{X}\times\{0,1\}$ and loss $\ell\colon \mathbb{R}^k\times\mathcal{Z}\to\{0,1\}$. Each $w \in \mathcal{W}$ corresponds to a classifier $f_w\colon \mathcal{X}\to\{0,1\}$ that can be interpreted as a deterministic neural network with parameters in $\mathbb{R}^k$. For trainable parameters $w_P \in \mathbb{R}^k$, $\gamma \in \mathbb{R}_+^k$, $\lambda \in \mathbb{R}_+$, let $\mathcal{G}$ be the set of all Gaussian posteriors of the form $P = \mathcal{N}(w_P, \mathrm{diag}(\gamma))$ and let $Q = \mathcal{N}(w_0, \lambda I_k)$ be a prior centered at a non-trainable random initialization, $w_0 \in \mathbb{R}^k$. We

can use a convex surrogate of the 0-1 loss, and the reparameterization trick $w = w_P + \nu \odot \sqrt{\gamma}$, $\nu \sim \mathcal{N}(0, I_k)$ (Kingma & Welling, 2014; Blundell et al., 2015) to compute an unbiased estimate of the gradient of the PAC-Bayes bound in Corollary 4 w.r.t. the parameters w_P, γ, λ and β. Computing the expectation $\mathbb{E}_P[L_S(f_W)]$ is difficult in practice. Instead, we use a Monte Carlo estimate $\hat{L}_S(f_W) = \frac{1}{m}\sum_{i=1}^m L_S(f_{W_i})$, where $W_i \overset{\text{i.i.d.}}{\sim} P$. Then, for any $\delta, \delta' \in (0,1)$, fixed $\alpha > 1$, $c \in (0,1)$, $b \in \mathbb{N}$, and $m, n \in \mathbb{N}$, with probability of at least $1 - \delta - \delta'$ over a draw of $S \sim \mu^{\otimes n}$ and $W \sim (P)^{\otimes m}$, we have

$$\mathbb{E}_P[L_\mu(f_W)] \leq \inf_{P \in \mathcal{G}, \beta > 1, \lambda \in (0,c)} \Phi_\beta^{-1}\Big\{\hat{L}_S(f_W) + \frac{\alpha}{n\beta} D(P\|Q) + R(\lambda, \beta; \delta, \delta')\Big\}, \tag{4}$$

where $R = \frac{2\alpha}{n\beta}\ln\Big[\frac{\ln \alpha^2 \beta n}{\ln \alpha}\Big] + \frac{\alpha}{n\beta}\ln\Big[\frac{\pi^2 b^2}{6\delta}\big(\ln\frac{c}{\lambda}\big)^2\Big] + \sqrt{\frac{1}{2m}\ln\frac{2}{\delta'}}$ accounts for the cost of optimizing the parameters β, λ, and using the Monte Carlo estimate of the empirical risk. For large n, m, R is negligible, and the optimization is dominated by the IC term, $\hat{L}_S(f_W) + \alpha(n\beta)^{-1} D(P\|Q)$.

Entropy-SGD. A related approach is *Entropy-SGD* (Chaudhari et al., 2017), which instead directly minimizes the stochastic complexity, $-\frac{1}{\beta}\ln \mathbb{E}_Q e^{-\beta L_S(W)}$. This entails optimizing the prior Q, when a PAC-Bayesian bound such as (1) stipulates that the prior be fixed before the draw of the training sample S. A way out is to sample Q in a differentially private fashion, and this forms the basis of the Entropy-SGLD algorithm (Dziugaite & Roy, 2018a). For $Q = \mathcal{N}\big(w, (\beta\gamma)^{-1} I_k\big)$, the stochastic complexity can be equivalently written as (up to constant terms) $-\frac{1}{\beta}\ln \int_{w' \in \mathbb{R}^k} e^{-\beta\left[L_S(w') + \frac{\gamma}{2}\|w - w'\|^2\right]} \mathrm{d}w'$, which can be interpreted as a measure of *flatness* of the loss landscape that measures the log-volume of low-loss parameter configurations around w. More generally, from the perspective of ICM, both Entropy- and PAC-Bayes- SGD can be viewed as optimization schemes that search for flat minima solutions (Hochreiter & Schmidhuber, 1997).

PAC-Bayes and Occam's factor. Lemma 6 gives the form of the optimal posterior under a quadratic approximation of the loss around a local minimizer:

Lemma 6. *Consider a quadratic approximation of the training loss around a local minimizer w_P, $\tilde{L}_S(w) = \frac{1}{2}(w - w_P)^\top H (w - w_P)$, a fixed prior $Q = \mathcal{N}(w_Q, \lambda^{-1} I_k)$, and a posterior distribution of the form $P = \mathcal{N}(w_P, \Sigma_P)$. Then the solution to the convex optimization problem $\min_{\Sigma_P} \mathbb{E}_P[\tilde{L}_S(W)] + (n\beta)^{-1} D(P\|Q)$, is given by $\Sigma_P^\star = H_\lambda^{-1}$, where $H_\lambda := (n\beta H + \lambda I_k)$. Here we assume $\lambda > 0$ is sufficiently large so that H_λ is positive definite.*

We can use a posterior of the form $P = \mathcal{N}(w_P, H_\lambda^{-1})$ to get the following PAC-Bayesian-style bound that incorporates second-order curvature information of the training loss:

Proposition 7. *Let $\{\lambda_i\}_{i=1}^k$ be the eigenvalues of H_λ and suppose that $\lambda_i \geq \lambda > 0$ for all i. Let $Q = \mathcal{N}(w_Q, \lambda^{-1} I_k)$ be a prior, and let $P = \mathcal{N}(w_P, H_\lambda^{-1})$. Then with probability of at least $1 - \delta$ over a draw of the sample S, we have*

$$\mathbb{E}_P[M_\beta(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\ln\frac{1}{\delta} + \frac{1}{n\beta}\Big(\frac{\lambda}{2}\|w_Q - w_P\|^2 + \frac{1}{2}\sum_{i=1}^k \ln\frac{\lambda_i}{\lambda}\Big). \tag{5}$$

The proof of Proposition 7 is given in Appendix A.6. Notably, the log-ratio term $\frac{1}{2}\sum_{i=1}^k \ln\frac{\lambda_i}{\lambda} = -\ln\sqrt{\det\frac{\lambda}{H_\lambda}}$ in (5) is the negative logarithm of the *Occam factor* (MacKay, 1992; Smith & Le, 2018). The log-Occam factor is the differential entropy of the Gaussian posterior with a scaled covariance $\lambda(H_\lambda)^{-1}$, and can be interpreted as the amount of information we gain about the model's parameters after seeing the training data. From the perspective of ICM, minimizing the right hand side of (5) w.r.t. the posterior leads to solutions with higher entropy and hence wider minima.

4 Discussion

We presented a unified treatment of PAC-Bayesian and information-theoretic generalization bounds, and obtained a few new results for data-dependent priors and unbounded losses. The bounds we studied are along the notion that bounded information (between the training data and the output hypothesis) implies learning. On the other hand, Bassily et al. (2018) and Nachum & Yehudayoff (2019) showed that learning does *not* imply bounded information. In particular, the information revealed by a learning algorithm about its input can be unbounded even for hypothesis classes of VC dimension 1. Livni & Moran (2020) showed a result in a similar vein for the PAC-Bayesian framework. Identifying the common structural properties of these negative results in the information-theoretic and PAC-Bayesian frameworks is an important avenue for further investigation.

ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757983).

REFERENCES

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Andrew R. Barron and Thomas M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 1046–1059, 2016.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 25–55, 2018.

Avrim Blum and John Langford. PAC-MDL bounds. In *Learning theory and kernel machines*, pp. 344–357. Springer, 2003.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1613–1622, 2015.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information based bounds on generalization error. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pp. 587–591. IEEE, 2019.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. Institute of Mathematical Statistics, 2007.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*, 2017.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pp. 2350–2358, 2015.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Gintare Karolina Dziugaite and Daniel M. Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1377–1386, 2018a.

Gintare Karolina Dziugaite and Daniel M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, pp. 8430–8441, 2018b.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pp. 535–544, 2018.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 353–360, 2009.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pp. 1884–1892, 2016.

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.

Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. *IEEE Transactions on Information Theory*, 56(1):438–449, 2009.

Marton Havasi, Robert Peharz, and Josè Miguel Hernàndez-Lobato. Minimal random code learning: Getting bits back from compressed model parameters. In *International Conference on Learning Representations*, 2019.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, pp. 824–839, 2020a.

Fredrik Hellström and Giuseppe Durisi. Generalization error bounds via mth central moments of the information density. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pp. 2741–2746. IEEE, 2020b.

Geoffrey E. Hinton and Drew van Camp. Keeping neural networks simple by minimising the description length of weights. In *Conference On Learning Theory*, pp. 5–13, 1993.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pp. 1475–1479. IEEE, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Ilja Kuzborskij, Nicolò Cesa-Bianchi, and Csaba Szepesvári. Distribution-dependent analysis of Gibbs-ERM principle. In *Conference on Learning Theory*, pp. 2028–2054, 2019.

John Langford and Rich Caruana. (Not) bounding the true error. In *Advances in Neural Information Processing Systems*, pp. 809–816, 2002.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018.

Ming Li, John Tromp, and Paul Vitányi. Sharpening Occam's razor. *Information Processing Letters*, 85(5):267–274, 2003.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, 1986.

Roi Livni and Shay Moran. A limitation of the PAC-Bayes framework. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ben London. A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 2931–2940, 2017.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

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

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pp. 164–170. ACM, 1999a.

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

David A. McAllester. A PAC-Bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *Journal of the ACM*, 63(3):1–10, 2016.

Ido Nachum and Amir Yehudayoff. Average-case information complexity of learning. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 633–646, 2019.

Jorma Rissanen. *Stochastic complexity in statistical inquiry*. World scientific, 1989.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–494. IEEE, 2016.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1232–1240, 2016.

Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3(Oct):233–269, 2002.

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

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Samuel L. Smith and Quoc V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference On Learning Theory*, pp. 3437–3452, 2020a.

Thomas Steinke and Lydia Zakynthinou. Open problem: Information complexity of VC learning. In *Conference on Learning Theory*, pp. 3857–3863, 2020b.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 466–492, 2017.

Tim van Erven. PAC-Bayes mini-tutorial: A continuous union bound. *arXiv preprint arXiv:1405.1580*, 2014.

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

Kenji Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory*, 44(4):1424–1439, 1998.

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

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

A APPENDIX

A.1 TECHNICAL LEMMAS

Proposition 8 collects some well-known facts about the cumulant generating function $\Lambda_X(\beta) = \ln \mathbb{E}[e^{\beta X}]$ of a random variable X for $\beta > 0$; see, e.g., (Boucheron et al., 2013, §2), and (Zhang, 2006a).

Proposition 8 (Facts about cumulant generating function)**.**

1. $\Lambda_X(\beta)$ *is convex in* β.
2. $\frac{1}{\beta}\Lambda_X(\beta)$ *is an increasing function of* β.
3. $\Lambda_{X-\mathbb{E}[X]}(0) = \Lambda'_{X-\mathbb{E}[X]}(0) = 0$.
4. *For real constants* a, b, $\Lambda_{aX+b}(\beta) = \Lambda_X(a\beta) + b$.
5. $\Lambda_X(\beta) \le \beta\mathbb{E}[X] + \frac{\beta^2}{2}\mathrm{Var}(X) + O(\beta^3)$, *where* $\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$.
6. *If* $X \in [0,1]$, *then* $\frac{1}{\beta}\Lambda_X(\beta) \le \frac{1}{\beta}\ln\big(1 - (1-e^{\beta})\mathbb{E}[X]\big)$, *with equality when* $X \in \{0,1\}$ *is Bernoulli.*
7. X *is* σ-sub-Gaussian *if* $\Lambda_{X-\mathbb{E}[X]}(\beta) \le \frac{\beta^2\sigma^2}{2}$.
8. X *is* (σ, c)-sub-gamma *if* $\Lambda_{X-\mathbb{E}[X]}(\beta) \le \frac{\beta^2\sigma^2}{2(1-c)}$ *for all* $\beta \in (0, \frac{1}{c})$.

Recall that the annealed expectation is $M_\beta(w) = -\beta^{-1}\ln \mathbb{E}_\mu[e^{-\beta\ell(w,Z)}] = -\beta^{-1}\Lambda_{-\ell(w,Z)}(\beta)$. By Proposition 8 item *1)* and Jensen's inequality, we have $M_\beta(w) \le L_\mu(w)$. For general loss functions, Proposition 8 item *5)* is useful for getting bounds in the opposite direction. By items *4)*, *7)* and *8)* of Proposition 8, if for all $w \in \mathcal{W}$, $\ell(w, Z)$ is σ-sub-Gaussian, resp., (σ, c)-sub-gamma under μ, then we have for all $w \in \mathcal{W}$ and $\beta > 0$, $L_\mu(w) \le M_\beta(w) + \frac{\beta}{2}\sigma^2$, resp., $L_\mu(w) \le M_\beta(w) + \frac{\beta}{2(1-c)}\sigma^2$.

We will need the following variational characterization of the KL divergence:

Lemma 9 (Donsker-Varadhan, Boucheron et al. 2013, Corollary 4.15)**.** *Let* P, Q *be probability measures on* $\mathcal{W}$, *and let* $\mathcal{F}$ *denote the set of measurable functions* $f : \mathcal{W} \to \mathbb{R}$ *such that* $\mathbb{E}_Q[e^{f(W)}] < \infty$. *If* $D(P\|Q) < \infty$, *then for every* $f \in \mathcal{F}$, *we have*

$$D(P\|Q) = \sup_{f\in\mathcal{F}}\Big\{\mathbb{E}_P[f(W)] - \ln \mathbb{E}_Q[e^{f(W)}]\Big\},$$

where the supremum is attained when $f = \ln\frac{\mathrm{d}P}{\mathrm{d}Q}$.

We note the following characterization of the inverse of the Fenchel-Legendre dual of a smooth convex function:

Lemma 10 (Boucheron et al. 2013, Lemma 2.4)**.** *Let* ψ *be a convex and continuously differentiable function defined on the interval* $[0, b)$, *where* $0 < b \le \infty$. *Assume that* $\psi(0) = \psi'(0) = 0$. *Then, the Legendre dual of* ψ, *defined as* $\psi^*(t) := \sup_{\beta\in[0,b)}\{\beta t - \psi(\beta)\}$, *is a nonnegative convex and nondecreasing function on* $[0,\infty)$ *with* $\psi^*(0) = 0$. *Moreover, for every* $y \ge 0$, *the set* $\{t \ge 0 : \psi^*(t) > y\}$ *is non-empty and the generalized inverse of* ψ^* *defined by* $\psi^{*-1}(y) = \inf\{t \ge 0 : \psi^*(t) > y\}$ *can also be written as* $\psi^{*-1}(y) = \inf_{\beta\in(0,b)}\frac{y+\psi(\beta)}{\beta}$.

We will need the following property of a Gibbs distribution:

Lemma 11 (Zhang 2006a, Proposition 3.1)**.** *For any real-valued measurable function* f *on* $\mathcal{W}$, *any real* $\beta > 0$, *and any* $P, Q \in \mathcal{M}(\mathcal{W})$ *such that* $D(P\|Q) < \infty$, *we have* $\beta^{-1}D(P\|P^*) = \mathbb{E}_P[f(W)] + \beta^{-1}D(P\|Q) + \beta^{-1}\ln\mathbb{E}_Q[e^{-\beta f(W)}]$, *where* P^* *is the Gibbs distribution* $P^*(\mathrm{d}w) := \frac{e^{-\beta f(w)}}{\mathbb{E}_Q[e^{-\beta f(W')}]}Q(\mathrm{d}w)$. *Consequently,*

$$\inf_{P\in\mathcal{M}(\mathcal{W})}\big\{\mathbb{E}_P f(W) + \beta^{-1}D(P\|Q)\big\} = -\beta^{-1}\ln\mathbb{E}_Q[e^{-\beta f(W)}].$$

Finally, we recall the *golden formula*: For all $Q \in \mathcal{M}(\mathcal{W})$ such that $D(P_W\|Q) < \infty$, we have

$$I(S;W) = D(P_{W|S}\|Q|P_S) - D(P_W\|Q), \tag{6}$$

where $D(P_{W|S}\|Q|P_S) = \int_{\mathcal{Z}^n} D(P_{W|S=s}\|Q)\mu^{\otimes n}(\mathrm{d}s)$.

By the golden formula (6), under the oracle prior $Q^\star = \mathbb{E}_S[P_{W|S}]$, $\mathbb{E}_S[D(P\|Q^\star)] = I(S;W)$.

All information-theoretic quantities are expressed in *nats*, unless specified otherwise.

A.2 Omitted details in Section 2

We include a proof of Lemma 1 since we will use the arguments.

Proof of Lemma 1. Applying the Donsker-Varadhan Lemma 9 to the function,

$$f(w) = n\beta(M_\beta(w) - L_S(w)), \tag{7}$$

we obtain,

$$n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q) \le \ln \mathbb{E}_Q[e^{n\beta(M_\beta(W) - L_S(W))}]. \tag{8}$$

Exponentiating both sides of (8) and taking expectations w.r.t. $S \sim \mu^{\otimes n}$, we have

$$\mathbb{E}_S \exp\left\{n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)\right\} \le \mathbb{E}_S\mathbb{E}_Q[e^{n\beta(M_\beta(W) - L_S(W))}]. \tag{9}$$

Since $Z_i \overset{\text{i.i.d.}}{\sim} \mu$, for any $w \in \mathcal{W}$ and $\beta > 0$, we have $e^{-n\beta M_\beta(w)} = \mathbb{E}_{S\sim\mu^{\otimes n}}\left[e^{-n\beta L_S(w)}\right]$. This observation and Fubini's theorem implies that the right hand side of (9) is equal to one. This proves the IEI. □

Proof of Theorem 2. Letting $R(S) := n\beta\mathbb{E}_P[M_\beta(W) - L_S(W)] - D(P\|Q)$, by Lemma 1, we have $\mathbb{E}_S[e^{R(S)}] \le 1$. By Markov's inequality,

$$\Pr_S\left(R(S) > \ln\frac{1}{\delta}\right) = \Pr_S\left(e^{R(S)} > \frac{1}{\delta}\right) \le \mathbb{E}_S[e^{R(S)}]\delta \le \delta.$$

Therefore, with probability of at least $1-\delta$ over the choice of $S \sim \mu^{\otimes n}$, we have for all $P \ll Q$,

$$\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta}\right). \tag{10}$$

By assumption,

$$\sup_{w\in\mathcal{W}}[L_\mu(w) - M_\beta(w)] \le \frac{\psi(\beta)}{\beta},\ \forall\beta > 0, \tag{11}$$

when (1) follows.

For the bound in expectation, note that by Jensen's inequality, $e^{\mathbb{E}_S[R(S)]} \le \mathbb{E}_S[e^{R(S)}] \le 1$, which implies $\mathbb{E}_S[R(S)] \le 0$, when we have

$$\mathbb{E}_{SW}[M_\beta(W)] \le \mathbb{E}_{SW}[L_S(W)] + \frac{1}{n\beta}D(P\|Q|P_S). \tag{12}$$

Using (11), and rearranging and optimizing, we have

$$\begin{aligned}\mathbb{E}_{SW}[\mathrm{g}(W,S)] &\le \inf_{\beta>0}\frac{\frac{1}{n}D(P\|Q|P_S) + \psi(\beta)}{\beta} \\ &= \psi^{*-1}\left(\frac{D(P\|Q|P_S)}{n}\right),\end{aligned}$$

where the second equality follows from Lemma 10. □

We can also optimize β in (1) at a small cost using the union bound:

Proposition 12. *Consider the setting in Theorem 2. If $\ell(w,Z)$ is σ-sub-Gaussian under μ for all $w \in \mathcal{W}$, then for any constants $\alpha > 1$ and $v > 0$, and any $\delta \in (0,1]$, for all $\beta \in (0,v]$, with probability of at least $1-\delta$, we have*

$$\mathbb{E}_P[\mathrm{g}(W,S)] \le \frac{\alpha}{n\beta}\left(D(P\|Q) + \ln\frac{\log_\alpha\sqrt{n} + K}{\delta}\right) + \frac{\beta\sigma^2}{2},$$

where $K = \max\{\log_\alpha\left(\frac{v\sigma}{\sqrt{2\alpha}}\right), 0\} + e$.

The proof follows that of van Erven (2014, Lemma 8), extending it to sub-Gaussian losses.

Proof. For $0 < u < v$, and $i = 0, \ldots, \lceil \log_\alpha \frac{v}{u} \rceil - 1$, for all i let $\beta_i = u\alpha^i$ be selected before the draw of the training sample. Then for every $\beta \in [u, v]$, there is a β_i such that $\beta_i \leq \beta \leq \alpha\beta_i$.

We can extend (10) by applying a union bound over the β_i's, so that for all P with probability of at least $1 - \delta$ over the draw of S, the following holds simultaneously for all β_i:

$$\mathbb{E}_P[M_{\beta_i}(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{\alpha}{n\beta_i}\left(D(P\|Q) + \ln \frac{\lceil \log_\alpha \frac{v}{u} \rceil}{\delta}\right). \tag{13}$$

By Proposition 8(2), for any $w \in \mathcal{W}$, $M_\beta(w)$ is a nonincreasing of β. Thus for any $\beta \in [u, v]$ and β_i such that $\beta_i \leq \beta \leq \alpha\beta_i$, $M_\beta(w) \leq M_{\beta_i}(w)$ and $\frac{1}{\beta_i} \leq \frac{\alpha}{\beta}$. Moreover, since $\ell(w, Z)$ is σ-sub-Gaussian under μ by assumption, we have for all $w \in \mathcal{W}$ and $\beta > 0$, $L_\mu(w) \leq M_\beta(w) + \frac{\beta}{2}\sigma^2$. Hence, with probability of at least $1 - \delta$ we have,

$$\mathbb{E}_P[L_\mu(W)] \leq \mathbb{E}_P[L_S(W)] + \frac{\alpha}{n\beta}\left(D(P\|Q) + \ln \frac{\lceil \log_\alpha \frac{v}{u} \rceil}{\delta}\right) + \frac{\beta\sigma^2}{2}. \tag{14}$$

Letting $J = D(P\|Q) + \ln \frac{\log_\alpha \sqrt{n}+K}{\delta}$, we find that the value for β that optimizes the right hand side of the bound in the statement of the proposition is bounded from below by $\sqrt{\frac{2\alpha}{n\sigma^2}}$. Letting $u = \frac{1}{\sqrt{n}} \min\left\{\sqrt{\frac{2\alpha}{\sigma^2}}, v\right\}$ and plugging it in (14) completes the proof. □

Remark 13 (Related work)**.** *A variation of the IEI for the special case of the 0-1 loss appears in the monograph by Catoni (Catoni, 2007, Eq. 1.2), and has been rediscovered more recently for the sub-Gaussian loss in (Hellström & Durisi, 2020b;a). The statements of (Hellström & Durisi, 2020b, Corollary 3, Eq. 20) and (Hellström & Durisi, 2020a, Corollary 2, Eq. 26) which are analogues of our Proposition 12 are incorrect as they assume that β can be optimized "for free," when in fact we have to pay a union bound price for optimizing β, which is selected* before *the draw of the training sample. We also note two related works that focus exclusively on unifying either PAC-Bayesian and Occam's razor-based bounds for the 0-1 loss (Blum & Langford, 2003), or information-theoretic bounds for the sub-Gaussian loss (Hafez-Kolahi et al., 2020).*

A.3 The Strong Functional Representation Lemma and Single-draw bounds

In this section, we highlight a functional characterization of the mutual information in relation to the "single-draw" bound due to Bassily et al. (2018).

A randomized learning algorithm $P_{W|S}$ can be viewed as a noisy channel that maps the input sample S to conditional distributions of hypotheses W in $\mathcal{W}$. Consider a communication task where Alice and Bob share a common random string R, possibly of unbounded length, generated in advance. Alice observes a sample $s \in \mathcal{Z}^n$ drawn according to P_S and communicates a prefix-free message M to Bob via a noiseless channel such that Bob can output a hypothesis $w \in \mathcal{W}$ that is distributed according to $P_{W|S=s}$. Harsha et al. (2009) showed that the minimum expected description length of M (in bits) needed to accomplish this task is roughly equal to the input-output mutual information $I(S; W)$. Variations on this theme have appeared in a learning-theoretic setting (Blum & Langford, 2003), and by way of the bits-back argument due to Hinton & van Camp (1993); see, e.g., Havasi et al. (2019). More generally, we note the following functional characterization of the mutual information:

Theorem 14 (Strong functional representation lemma (SFRL), Li & El Gamal 2018)**.** *For any pair of jointly distributed random variables (S, W) with $I(S; W) < \infty$, there exists a random variable R independent of S such that W can be represented as a deterministic function of S and R, and*

$$I(S; W) \leq H(W|R) \leq I(S; W) + \log(I(S; W) + 1) + 4.$$

The SFRL implies the existence of a random variable $R \perp\!\!\!\perp S$ such that $H(W|R) \approx I(S; W)$.

Consider the case for the $\{0, 1\}$-valued loss. If the algorithm $P_{W|S}$ is deterministic, then we have $I := I(S; W) = H(W) - H(W|S) = H(W)$. By Markov's inequality, with probability of at least $1 - \delta$, we have $P_W(w) \geq e^{-I/\delta}$. Let $\mathcal{W}_0 \subseteq \mathcal{W}$ be the set of hypotheses so that $P_W(w) \geq e^{-I/\delta}$. The size of $\mathcal{W}_0$ is at most $e^{I/\delta}$ since $1 = \Pr(\mathcal{W}) \geq \Pr(\mathcal{W}_0) = \sum_{w \in \mathcal{W}_0} P_W(w) \geq |\mathcal{W}_0| e^{-I/\delta}$. By the Chernoff-Hoeffding bound, for every w in $\mathcal{W}$,

$$\Pr_S\Big(|\mathrm{g}(w, S)| > \epsilon\Big) \leq 2e^{-2n\epsilon^2} \quad \forall \epsilon > 0.$$

Applying the union bound over all $w \in \mathcal{W}_0$, the probability of error for the algorithm is $2|\mathcal{W}_0|e^{-2n\epsilon^2} + \delta$, where the second summand is for the case where the algorithm outputs a function outside $\mathcal{W}_0$. Hence, for every $w \in \mathcal{W}_0$, the empirical risk is close to the true risk for $n = \Omega(\frac{I}{\delta\epsilon^2})$ with probability of at least $1 - \delta$.

By the SFRL, any randomized algorithm can be simulated by randomly sampling a deterministic algorithm from some distribution R before observing the input S. These algorithms have the property that on average (over R), $H(W|R) \approx I(S;W)$. Using the argument for the deterministic case and integrating over R, we can bound the probability of error for the randomized case as

$$\Pr_{S,W}\Big(|\mathrm{g}(W,S)| > \epsilon\Big) = O\bigg(\frac{I(S;W)}{n\epsilon^2}\bigg). \tag{15}$$

An analogous bound for the sub-Gaussian loss appears in (Xu & Raginsky, 2017, Theorem 3).

A.4 Recovering classical PAC-Bayesian bounds

By Proposition 8 item *6)*, for a $\{0,1\}$-valued loss, we have

$$M_\beta(w) = -\beta^{-1}\ln\big(1 - (1 - e^{-\beta})L_\mu(w)\big) =: \Phi_\beta(L_\mu(w)).$$

Φ_β is an increasing one-to-one mapping of the unit interval onto itself, and is convex for $\beta > 0$ (Catoni, 2007). The inverse of Φ_β is given by $\Phi_\beta^{-1}(x) = \frac{1-e^{-\beta x}}{1-e^{-\beta}}$, and we recover Catoni's PAC-Bayesian bound in Corollary 4.

Starting from Catoni's bound in Corollary 4, using $1 \le \beta(1-e^{-\beta})^{-1} \le (1-\frac{\beta}{2})^{-1}$, we have

$$\begin{aligned}
\mathbb{E}_P[L_\mu(W)] &\le \Phi_\beta^{-1}\Big\{\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\Big[D(P\|Q) + \ln\frac{1}{\delta}\Big]\Big\} \\
&\le \frac{\beta}{1-e^{-\beta}}\bigg[\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\bigg(D(P\|Q) + \ln\frac{1}{\delta}\bigg)\bigg] \qquad (16)\\
&\le \frac{1}{1-\frac{\beta}{2}}\bigg[\mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\bigg(D(P\|Q) + \ln\frac{1}{\delta}\bigg)\bigg]. \qquad (17)
\end{aligned}$$

(16) and (17) recover, resp., the bounds due to Catoni (2007, Theorem 1.2.1) and McAllester (2013, Theorem 2), where for the latter we additionally require that $\beta < 2$.

We now show how inequality (10) relates to other well-known PAC-Bayesian inequalities such as the "PAC-Bayes-KL-inequality" (Seeger, 2002; Maurer, 2004). Applying the Donsker-Varadhan lemma to the function $f(w) = n\beta(L_\mu(w) - L_S(w))$, which involves the true risk $L_\mu(w)$ instead of the annealed expectation $M_\beta(w)$ (see 7), and following the same steps as in the proof of (10) in Theorem 2, we arrive at the following PAC-Bayesian bound:

$$\Pr_{S\sim\mu^{\otimes n}}\Bigg(\mathbb{E}_P[L_\mu(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\Big[D(P\|Q) + \ln\frac{1}{\delta} + \ln\mathbb{E}_Q\mathbb{E}_{S'\sim\mu^{\otimes n}}\, e^{n\beta\left(L_\mu(W) - L_{S'}(W)\right)}\Big]\Bigg) \ge 1-\delta. \tag{18}$$

For an explicit comparison of (18) with (10), we write the latter as

$$\Pr_{S\sim\mu^{\otimes n}}\Bigg(\mathbb{E}_P[M_\beta(W)] \le \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\Big[D(P\|Q) + \ln\frac{1}{\delta} + \underbrace{\ln\mathbb{E}_Q\mathbb{E}_{S'\sim\mu^{\otimes n}}\, e^{n\beta\left(M_\beta(W) - L_{S'}(W)\right)}}_{=0}\Big]\Bigg) \ge 1-\delta, \tag{19}$$

where the last term in the right hand side of the bound in (19) vanishes since

$$e^{-n\beta M_\beta(w)} = \mathbb{E}_{S'\sim\mu^{\otimes n}}\Big[e^{-n\beta L_{S'}(w)}\Big]$$

for any $w \in \mathcal{W}$ and $\beta > 0$. In contrast, the term $\ln\mathbb{E}_Q\mathbb{E}_{S'\sim\mu^{\otimes n}}\, e^{n\beta(L_\mu(W) - L_{S'}(W))}$ involving the true risk in (18) is, in general, positive.

Specializing to the case of a $\{0,1\}$-valued loss, fix $\beta = 1$, and let $\Delta : [0,1] \times [0,1] \to \mathbb{R}$ be a convex function. Applying the Donsker-Varadhan lemma to the function, $f(w) = n\Delta(L_S(w), L_\mu(w))$, following the same steps as in the proof of (10) in Theorem 2, and by noting that $\Delta(\mathbb{E}_P[L_S(W)], \mathbb{E}_P[L_\mu(W)])) \leq \mathbb{E}_P[\Delta(L_S(W), L_\mu(W))]$, we arrive at the following PAC-Bayesian bound (see, e.g., Maurer (2004, Lemma 3), Germain et al. (2009, Theorem 2.1), Rivasplata et al. (2020, Eq. 4)):

$$\Pr_{S\sim\mu^{\otimes n}}\left(\Delta(\mathbb{E}_P[L_S(W)], \mathbb{E}_P[L_\mu(W)])) \leq \frac{1}{n}\Big[D(P\|Q) + \ln\frac{1}{\delta} + \ln \mathbb{E}_Q\mathbb{E}_{S'\sim\mu^{\otimes n}}\, e^{n\Delta(L_{S'}(W), L_\mu(W))}\Big]\right) \geq 1-\delta. \tag{20}$$

For $x, y \in [0,1]$, the binary KL divergence is $\mathrm{kl}(y\|x) = y\ln\frac{y}{x} + (1-y)\ln\frac{1-y}{1-x}$. The PAC-Bayes-KL-inequality comes about by upper-bounding the log-exponential-moment term involving the true risk in the right hand side of the bound in (20): For $\Delta(y,x) = \mathrm{kl}(y\|x)$, Maurer (2004) showed that for $n \geq 8$, $\mathbb{E}_Q\mathbb{E}_{S'}[e^{n\Delta(L_{S'}(W), L_\mu(W))}] \leq 2\sqrt{n}$, when we have

$$\Pr_{S\sim\mu^{\otimes n}}\left(\mathrm{kl}(\mathbb{E}_P[L_S(W)], \mathbb{E}_P[L_\mu(W)]) \leq \frac{1}{n}\Big[D(P\|Q) + \ln\frac{2\sqrt{n}}{\delta}\Big]\right) \geq 1-\delta. \tag{21}$$

Letting $\Delta(y,x) = 2(y-x)^2$ leads to the bound in (McAllester, 1999b), while letting $\Delta(y,x) = (y-x)^2/(2x)$ leads to that in (Thiemann et al., 2017).

Fixing $\beta = 1$ in (1), we recover (Germain et al., 2016, Corollary 5):

Corollary 15. *Consider the setting in Theorem 2. If the loss ℓ is (σ, c)-sub-gamma with $c < 1$, then with probability of at least $1-\delta$ over the choice of $S \sim \mu^{\otimes n}$, for all distributions $P \ll Q$ over $\mathcal{W}$,* $\mathbb{E}_P[\mathrm{g}(W,S)] \leq \frac{1}{n}(D(P\|Q) + \ln(1/\delta)) + \frac{\sigma^2}{2(1-c)}$.

The condition $c < 1$ guarantees that $\beta = 1 \in (0, \frac{1}{c})$ when the sub-gamma condition in Proposition 8 *8)* is satisfied. In the limit $c \to 0_+$, a sub-gamma loss reduces to the sub-Gaussian loss (Boucheron et al., 2013, §2.4), and we recover (Germain et al., 2016, Corollary 4).

A.5 Proof of Proposition 5

For the proof of Proposition 5, the key quantity of interest is the approximate max-information between the input S and the data-dependent prior. We note the following definitions.

For $\alpha \geq 0$, the *α-approximate max-divergence* is defined as

$$D_\infty^\alpha(P\|Q) = \ln \sup_{\mathcal{O}\subseteq\mathcal{X}:\, P(\mathcal{O})>\alpha} \frac{P(\mathcal{O}) - \alpha}{Q(\mathcal{O})}.$$

The *max-divergence* $D_\infty(P\|Q)$ is defined as $D_\infty^\alpha(P\|Q)$ for $\alpha = 0$. For a pair of variables (X, Y) with joint law P_{XY} and marginals P_X and P_Y, the *α-approximate max-information* between X and Y is defined as

$$I_\infty^\alpha(X;Y) = D_\infty^\alpha(P_{XY}\|P_X \otimes P_Y).$$

The *max-information* $I_\infty(X;Y)$ is defined to be $I_\infty^\alpha(X;Y)$ for $\alpha = 0$. $I_\infty(X;Y)$ is an upper bound on the ordinary mutual information $I(X;Y)$ (Dwork et al., 2015).

Definition 16 (Differential Privacy, Dwork & Roth 2014)**.** *For any $\epsilon \geq 0$ and $\delta \in [0,1]$, a learning algorithm $P_{W|S}$ is said to be* (ϵ, δ)-differentially private *if for all pairs of datasets $s, s' \in \mathcal{Z}^n$ that differ in a single element, $D_\infty^\delta(P_{W|S=s}\|P_{W|S=s'}) \leq \epsilon$. The case $\delta = 0$ is called pure differential privacy.*

Definition 17 (Max-Information of an algorithm, Dwork et al. 2015)**.** *We say that an algorithm $P_{W|S}$ has α-approximate max-information of k, denoted as $I_{\infty,\mu}^\alpha(P_{W|S}, n) \leq k$, if for every distribution μ over $\mathcal{Z}$, we have $I_\infty^\alpha(S;W) \leq k$ when $S \sim \mu^{\otimes n}$.*

It follows from the definition of α-approximate max-information that if an algorithm $P_{W|S}$ has bounded approximate max-information, then we can control the probability of "bad events" that may

arise as a result of the dependence of the output W on the input S (Dwork et al., 2015). Let $S' \perp\!\!\!\perp W$ be an independent sample with the same distribution as S. If for some $\alpha \geq 0$, $I^{\alpha}_{\infty}(S;W) \leq k$, then for any event $\mathcal{O} \subseteq \mathcal{Z}^n \times \mathcal{W}$, we have

$$\Pr((S,W) \in \mathcal{O}) \leq e^k \cdot \Pr((S',W) \in \mathcal{O}) + \alpha. \tag{22}$$

Pure differential privacy implies a bound on the approximate max-information:

Theorem 18 (Pure differential privacy and α-approximate max-information, Dwork et al. 2015, Theorem 20)**.** *If $P_{W|S}$ is an $(\epsilon, 0)$-differentially private algorithm, then $I_{\infty,\mu}(P_{W|S}, n) \leq \epsilon n$, and for any $\alpha > 0$, $I^{\alpha}_{\infty,\mu}(P_{W|S}, n) \leq n\epsilon^2/2 + \epsilon\sqrt{n\ln(2/\alpha)/2}$.*

Remark 19. *The result above is extended to (ϵ, δ)-differential privacy in (Rogers et al., 2016, Theorem 3.1): If $P_{W|S}$ is an (ϵ, δ)-differentially private algorithm for $\epsilon \in (0, 1/2]$ and $\delta \in (0, \epsilon)$, then for $\alpha = O(n\sqrt{\delta/\epsilon})$, $I^{\alpha}_{\infty,\mu}(\mathcal{A}, n) = O(n\epsilon^2 + n\sqrt{\delta/\epsilon})$.*

The proof of Proposition 5 follows closely that of (Dziugaite & Roy, 2018b, Theorem 4.2).

Proof of Proposition 5. For every $Q \in \mathcal{M}(\mathcal{W})$, let

$$F(Q) = \left\{ S' \in \mathcal{Z}^n \colon \exists\, P \in \mathcal{M}(\mathcal{W}),\, \mathbb{E}_P[M_\beta(W)] \geq \mathbb{E}_P[L_{S'}(W)] + \frac{1}{n\beta}\left(D(P\|Q) + \ln\frac{1}{\delta'} \right) \right\}.$$

By (10), we have $\Pr_{S' \sim \mu^{\otimes n}}\left(S' \in F(Q)\right) \leq \delta'$. From (22), we have

$$\Pr_{S\sim\mu^{\otimes n}}\left(S \in F(Q^0(S))\right) \leq e^{I^{\alpha}_{\infty,\mu}(Q^0,n)} \cdot \Pr_{(S,S')\sim\mu^{\otimes 2n}}\left(S' \in F(Q^0(S))\right) + \alpha \leq e^{I^{\alpha}_{\infty,\mu}(Q^0,n)} \cdot \delta' + \alpha.$$

Letting $\delta := e^{I^{\alpha}_{\infty,\mu}(Q^0,n)} \cdot \delta' + \alpha$, for $\alpha \in (0, \delta)$ we have,

$$\Pr_{S\sim\mu^{\otimes n}}\Bigg(\exists\, P \in \mathcal{M}(\mathcal{W}),\, \mathbb{E}_P[M_\beta(W)] \geq \mathbb{E}_P[L_S(W)] + \frac{1}{n\beta}\left(D(P\|Q^0(S)) + \ln\frac{1}{\delta-\alpha} + I^{\alpha}_{\infty,\mu}(Q^0,n) \right) \Bigg) \leq \delta.$$

The proof is complete by supplanting $I^{\alpha}_{\infty,\mu}(Q^0,n)$ with the bound in Theorem 18, and choosing $\alpha = \frac{\delta}{2}$. □

By Remark 19, Proposition 5 can be extended to (ϵ, δ)-differentially private priors.

We can also bound the expected generalization error:

Proposition 20. *Consider the setting in Theorem 2. Let $Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{W})$ be an $(\epsilon, 0)$-differentially private algorithm. Then with probability of at least $1 - \delta$ over a draw of the sample S, for all P, we have*

$$\mathbb{E}_P[\mathrm{g}(W,S)] \leq \frac{1}{n\beta}\left(D(P\|Q^0(S)) + \ln\frac{2}{\delta} + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\ln\frac{4}{\delta}} \right) + \frac{\psi(\beta)}{\beta}. \tag{23}$$

The main advantage of the max-information formulation in Propositions 5 and 20 is that we can get high probability guarantees at the cost of a $O(n\epsilon^2 + \epsilon\sqrt{n\ln 1/\delta})$ correction term. This cost is compensated for by a lower KL complexity since the prior is more "aligned" with the data-dependent posterior than when chosen independently of the data. As is well-known (Dwork et al., 2015; Feldman & Steinke, 2018), a small mutual information between the data and the prior will not ensure that bad events will happen with low probability.

A.6 Proofs for Section 3

When $\mathcal{G} = \mathcal{M}(\mathcal{W})$, applying Lemma 11 to $f(w) = nL_S(w)$, and writing β for $n\beta$, we recover the Gibbs algorithm, $P^\star$, in which case $\mathrm{OIC}^{\beta}_{\mathcal{G}}$ in (3) evaluates to the *(extended) stochastic complexity*, $-\frac{1}{\beta}\ln\mathbb{E}_Q[e^{-\beta L_S(W)}]$.

We show how to optimize the bound in Corollary 4 w.r.t. the parameters β and λ.

Proof of (4). First, note that the bound in Corollary 4 holds uniformly for all $\beta > 1$ at an additional cost arising from a union bound argument (Catoni, 2007, Theorem 1.2.7): For $\alpha > 1$,

$$\mathbb{E}_P[L_\mu(f_W)] \leq \inf_{\beta>1} \Phi_\beta^{-1}\Big\{\mathbb{E}_P[L_S(f_W)] + \frac{\alpha}{n\beta}\Big[D(P\|Q) + \ln\frac{1}{\delta} + 2\ln\Big(\frac{\ln\alpha^2\beta n}{\ln\alpha}\Big)\Big]\Big\}. \tag{24}$$

Second, we select λ before the draw of the training sample from a finite grid of possible values: Following (Langford & Caruana, 2002; Dziugaite & Roy, 2017), let $\lambda = ce^{-j/b}$ for some $j \in \mathbb{N}$ and fixed $b \in \mathbb{N}$, $c \in (0,1)$, where b and c control, resp., the resolution and size of the grid. If (24) holds for each $j \in \mathbb{N}$ with probability of at least $1 - \frac{6\delta}{\pi^2 j^2}$, then by the union bound, it holds for all $j \in \mathbb{N}$ simultaneously with probability of at least $1-\delta$, since $\sum_{j=1}^\infty \frac{6}{\pi^2 j^2} = 1$. Solving for j in terms of λ, we have

$$\mathbb{E}_P[L_\mu(f_W)] \leq \inf_{\beta>1,\,\lambda\in(0,c)} \Phi_\beta^{-1}\Bigg\{\mathbb{E}_P[L_S(f_W)] + \frac{\alpha}{n\beta}\Bigg[D(P\|Q) + \ln\Bigg(\frac{\pi^2 b^2}{6\delta}\Big(\ln\frac{c}{\lambda}\Big)^2\Bigg) + 2\ln\Big(\frac{\ln\alpha^2\beta n}{\ln\alpha}\Big)\Bigg]\Bigg\}.$$

Finally, we account for the cost of using a Monte Carlo estimate of the empirical risk, $\hat{L}_S(f_W) = \frac{1}{m}\sum_{i=1}^m L_S(f_{W_i})$, where $W_i \overset{\text{i.i.d.}}{\sim} P$. By an application of the Chernoff bound (Langford & Caruana, 2002, Theorem 2.5) and Pinsker's inequality, for any $\delta' \in (0,1)$, we have with probability of at least $1-\delta'$, $\mathbb{E}_P[L_S(f_W)] \leq \hat{L}_S(f_W) + \sqrt{\frac{1}{2m}\ln\frac{2}{\delta'}}$.

By another application of the union bound, Corollary 4 finally takes the form: For any $\delta, \delta' \in (0,1)$, fixed $\alpha > 1$, $c \in (0,1)$, $b \in \mathbb{N}$, and $m, n \in \mathbb{N}$, with probability of at least $1-\delta-\delta'$ over a draw of $S \sim \mu^{\otimes n}$ and $W \sim (P)^{\otimes m}$,

$$\mathbb{E}_P[L_\mu(f_W)] \leq \inf_{P\in\mathcal{G},\beta>1,\lambda\in(0,c)} \Phi_\beta^{-1}\Big\{\hat{L}_S(f_W) + \frac{\alpha}{n\beta}D(P\|Q) + R(\lambda,\beta;\delta,\delta')\Big\},$$

where $R(\lambda,\beta;\delta,\delta') = \frac{2\alpha}{n\beta}\ln\Big[\frac{\ln\alpha^2\beta n}{\ln\alpha}\Big] + \frac{\alpha}{n\beta}\ln\Big[\frac{\pi^2 b^2}{6\delta}\big(\ln\frac{c}{\lambda}\big)^2\Big] + \sqrt{\frac{1}{2m}\ln\frac{2}{\delta'}}$. □

Proof of Lemma 6. Letting $\theta = w - w_P$, and $P' = P - w_P$, note that $\theta^\top H\theta = \operatorname{Tr}(\theta^\top H\theta) = \operatorname{Tr}(H\theta\theta^\top)$. Hence

$$\mathbb{E}_{P'}[\tfrac{1}{2}\theta^\top H\theta] = \mathbb{E}_{P'}[\tfrac{1}{2}\operatorname{Tr}(H\theta\theta^\top)] = \tfrac{1}{2}\operatorname{Tr}(H\mathbb{E}_{P'}[\theta\theta^\top]) = \tfrac{1}{2}\operatorname{Tr}(H\Sigma_P).$$

For $Q \sim \mathcal{N}(w_Q, \Sigma_Q)$ and $P \sim \mathcal{N}(w_P, \Sigma_P)$, we have

$$\begin{aligned}
&\mathbb{E}_{P'}[\tfrac{1}{2}\theta^\top H\theta] + (n\beta)^{-1}D(P\|Q) \\
&\quad= \tfrac{1}{2}\operatorname{Tr}(H\Sigma_P) + (n\beta)^{-1}D(P\|Q) \\
&\quad= \frac{\operatorname{Tr}(H\Sigma_P)}{2} + \frac{(n\beta)^{-1}}{2}\Big(\ln\frac{\det\Sigma_Q}{\det\Sigma_P} + \operatorname{Tr}\Big(\Sigma_Q^{-1}\Sigma_P\Big) - k + (w_Q - w_P)^\top\Sigma_Q^{-1}(w_Q - w_P)\Big).
\end{aligned}$$

The derivative of the RHS w.r.t. Σ_P is $\frac{1}{2}\Big[H - (n\beta)^{-1}\Sigma_P^{-1} + (n\beta)^{-1}\Sigma_Q^{-1}\Big]^\top$, where we have used the fact that $\nabla_A \operatorname{Tr}(AB) = B^\top$, and $\nabla_A \ln\det(A) = (A^{-1})^\top$. Setting the derivative to zero and $\Sigma_Q = \lambda^{-1}I_k$ yields the result. □

Proof of Proposition 7. The proof follows from (10), and the fact that for $Q = \mathcal{N}(w_Q, \lambda^{-1}I_k)$, $P = \mathcal{N}(w_P, H_\lambda^{-1})$ such that $\lambda_i \geq \lambda > 0$ for all i, we have

$$D(P\|Q) = \frac{1}{2}\Bigg(\lambda\|w_Q - w_P\|^2 + \sum_{i=1}^k \ln\frac{\lambda_i}{\lambda} + \sum_{i=1}^k\Big(\frac{\lambda}{\lambda_i} - 1\Big)\Bigg) \leq \frac{1}{2}\Bigg(\lambda\|w_Q - w_P\|^2 + \sum_{i=1}^k \ln\frac{\lambda_i}{\lambda}\Bigg).$$

□