PAC-Bayes Analysis with Stochastic Kernels

Omar Rivasplata
University College London & DeepMind
o.rivasplata@cs.ucl.ac.uk

Ilja Kuzborskij
DeepMind
iljak@google.com

Csaba Szepesvári
DeepMind
szepi@google.com

John Shawe-Taylor
University College London
jst@cs.ucl.ac.uk

Abstract

We focus on a stochastic learning model where the learner observes a finite set of training examples and the output of the learning process is a data-dependent distribution over a space of hypotheses. The learned data-dependent distribution is then used to make randomized predictions, and the high-level theme addressed here is guaranteeing the learner's performance on examples that were not seen during training, i.e. generalization. In this setting the unknown quantity of interest is the expected risk of the data-dependent randomized predictor, for which upper bounds can be derived via a PAC-Bayes analysis, leading to PAC-Bayes bounds.

Specifically, we present a general form of the PAC-Bayes inequality, from which one may derive extensions of various known PAC-Bayes bounds as well as novel bounds. We clarify the role of the requirement of fixed 'data-free' priors and discuss the use of data-dependent priors. We also discuss a simple PAC-Bayes bound that is valid for loss functions with unbounded range. Our analysis clarifies that those two requirements are used to bound an exponential moment, while the general PAC-Bayes inequality remains valid with those restrictions removed.

1 Introduction

The context of this paper is the statistical learning model where the learner observes training data $S = (Z_1, Z_2, \ldots, Z_n)$ randomly drawn from a space of size-n samples $\mathcal{S} = \mathcal{Z}^n$ (e.g. $\mathcal{Z} = \mathbb{R}^d \times \mathcal{Y}$) according to some unknown probability distribution[1] $P_n \in \mathcal{M}_1(\mathcal{S})$. Typically $Z_1, \ldots, Z_n$ are independent and share a common distribution $P_1 \in \mathcal{M}_1(\mathcal{Z})$. Upon observing the training data S, the learner outputs a *data-dependent* probability distribution Q_S over a *hypothesis space* $\mathcal{H}$. Notice that this learning scenario involves randomness in the data and the hypothesis. In this stochastic learning model, the randomized predictions are carried out by randomly drawing a fresh hypothesis for each prediction. Therefore, we consider the performance of a probability distribution Q over the hypothesis space: the expected *population loss* is $Q[L] = \int_{\mathcal{H}} L(h)Q(dh)$, i.e. the Q-average of the standard population loss $L(h) = \int \ell(h, z)P_1(dx)$ for a fixed hypothesis $h \in \mathcal{H}$, where $\ell : \mathcal{H} \times \mathcal{Z} \to [0, \infty)$ is a given loss function and $P_1 \in \mathcal{M}_1(\mathcal{Z})$ generates one random example. Similarly, the expected *empirical loss* is $Q[\hat{L}_S] = \int_{\mathcal{H}} \hat{L}_S(h)Q(dh)$, where $\hat{L}_s(h) = \hat{L}(h, s)$ is the empirical loss, namely, $\hat{L}(h, s) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i)$ for a fixed h and $s = (z_1, \ldots, z_n)$.

An important component of our development is using a convenient way to formalize the notion of "data-dependent distributions over $\mathcal{H}$" that makes explicit their difference to fixed distributions.

[1] We write $\mathcal{M}_1(\mathcal{Z})$ to denote the family of probability measures over a set $\mathcal{Z}$, see Appendix A.

NeurIPS 2019 Workshop on Machine Learning with Guarantees, Vancouver, Canada.

Randomised predictors with a data-dependent distribution. A data-dependent distribution over the space $\mathcal{H}$ is formalized here as a *stochastic kernel*[2] defined as a mapping[3] $Q : \mathcal{S} \times \Sigma_{\mathcal{H}} \to [0, 1]$ such that (i) for each $B \in \Sigma_{\mathcal{H}}$ the function $s \mapsto Q(s, B)$ is measurable; and (ii) for each $s \in \mathcal{S}$ the function $B \mapsto Q(s, B)$ is a probability measure over $\mathcal{H}$. We will write $\mathcal{K}(\mathcal{S}, \mathcal{H})$ to denote the set of all such stochastic kernels from $\mathcal{S}$ to —distributions over— $\mathcal{H}$. In the following, given $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ and $s \in \mathcal{S}$, we will write $Q_s[L] = \int L(h)Q_s(dh)$ and $Q_s[\hat{L}_s] = \int \hat{L}_s(h)Q_s(dh)$ to denote the expected population loss and expected empirical loss, respectively.

With the notation just introduced, Q_S stands for the distribution over $\mathcal{H}$ corresponding to a randomly drawn data set S. The stochastic kernel Q can be thought of as describing a randomizing learner. One well-known example is the *Gibbs* learner, where Q_S is of the form $Q_S(dh) \propto e^{-\gamma \hat{L}(h,S)}\mu(dh)$ for some $\gamma > 0$, with μ a base measure over $\mathcal{H}$.

A common question arising in learning theory aims to explain the generalization ability of a learner: how can a learner ensure a 'well-behaved' population loss? One way to answer this question is via upper bounds on the population loss, also called *generalization bounds*. Often the focus is on the *generalization gap*, which is the difference between the population loss and the empirical loss, and giving upper bounds on the gap. There are several types of generalization bounds we care about in learning theory, with variations in the way they depend on the training data S and the data-generating distribution P_n. The classical bounds (such as VC-bounds) depend on neither. *Distribution-dependent* bounds are expressed in terms of quantities related to the data-generating distribution (e.g. population mean or variance) and possibly constants, but not the data in any way. These bounds can be helpful to predict the behaviour of a learning method on different distributions—for example, some data-generating distributions might give faster convergence rates than others. Finally, there are *data-dependent* bounds which are expressed in terms of empirical quantities that can be computed directly from data. These are of interest in practical situations, for instance for "self-bounding" algorithms, which are learning algorithms that use all the data to simultaneously provide a predictor and a certificate of performance [Freund, 1998].

PAC-Bayesian inequalities allow to derive distribution- or data-dependent generalization bounds in the context of the stochastic prediction model discussed above. The usual PAC-Bayes analysis introduces a reference probability measure $Q^0 \in \mathcal{M}_1(\mathcal{H})$ on the hypothesis space $\mathcal{H}$. The learned data-dependent distribution Q_S is commonly called a *posterior*, while Q^0 is called a *prior*. However, in contrast to Bayesian inference, the PAC-Bayes prior Q^0 acts as an analytical device and may or may not be used by the learning algorithm, and the PAC-Bayes posterior Q_S is unrestricted and may be different from the posterior that would be obtained from Q^0 through Bayesian inference.

2 Our Contributions

In this paper we discuss a general PAC-Bayesian theorem encompassing many usual bounds which appear in the literature [McAllester, 1998, Seeger, 2002, Catoni, 2007, Thiemann et al., 2017], but the formulation discussed here (see Theorem 2 in Appendix B) allows the PAC-Bayes priors to be data-dependent by default, and the loss functions to have an unbounded range.

Our take on the PAC-Bayes theorem (Theorem 2 in Appendix B) establishes that for any convex function $F : \mathbb{R}^2 \to \mathbb{R}$, probability kernels $Q, Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ and $\delta \in (0, 1)$,

$$F(Q_S[L], Q_S[\hat{L}_S]) \leq \mathrm{KL}(Q_S \| Q^0_S) + \log(\xi(Q^0)/\delta) \qquad \text{w.p.} \geq 1 - \delta \,, \tag{1}$$

where KL stands for the Kullback-Leibler divergence[4], and $\xi(Q^0)$ is the exponential moment of $F(L(h), \hat{L}_S(h))$, which is defined as follows:

$$\xi(Q^0) = \int_{\mathcal{S}} \int_{\mathcal{H}} e^{F(L(h), \hat{L}_S(h))} Q^0_s(dh) P_n(ds) \,.$$

[2] This is also called a transition kernel, see e.g. Kallenberg [2017] for more details on this definition.

[3] The space of size-n samples $\mathcal{S}$ is equipped with a sigma algebra that we denote $\Sigma_{\mathcal{S}}$, and the hypothesis space $\mathcal{H}$ is equipped with a sigma algebra $\Sigma_{\mathcal{H}}$. For precise definitions see Appendix A.

[4] Given two probability distributions $Q, Q' \in \mathcal{M}_1(\mathcal{H})$, the Kullback-Leibler divergence between them, also known as relative entropy, is defined as follows: $\mathrm{KL}(Q\|Q') = \int_{\mathcal{H}} \log\left(dQ/dQ'\right) dQ$, where dQ/dQ' denotes the Radon-Nikodym derivative. For Bernoulli distributions with parameters q and q' we will write $\mathrm{kl}(q\|q') = q \log(\frac{q}{q'}) + (1 - q) \log(\frac{1-q}{1-q'})$, also called the binary KL divergence.

Observe that Eq. (1) is defined for an arbitrary[5] convex function F. This way the usual bounds are encompassed. For example, taking $F(x,y) = 2n(x-y)^2$ yields the McAllester [1998]-type bound, $F(x,y) = n\,\mathrm{kl}(y\|x)$ gives the bound of Seeger [2002], by $F(x,y) = n\log\left(\frac{1}{1-x(1-e^{-\lambda})}\right) - \lambda n y$ we get the bound of Catoni [2007], whereas by $F(x,y) = n(x-y)^2/(2x)$ we obtain with a suitable derivation a bound of Thiemann et al. [2017], or by a different derivation we get a novel bound that holds under the usual requirements of fixed 'data-free' prior and losses within the $[0,1]$ range:

$$Q_S[L] \le \left(\sqrt{Q_S[\hat{L}_S] + \frac{\mathrm{KL}(Q_S\|Q^0) + \log(\frac{2\sqrt{n}}{\delta})}{2n}} + \sqrt{\frac{\mathrm{KL}(Q_S\|Q^0) + \log(\frac{2\sqrt{n}}{\delta})}{2n}}\right)^2 . \tag{2}$$

As consequence of the universality of Eq. (1), besides the usual bounds we may derive novel bounds, e.g. with data-dependent priors Q^0_S. Conceptually, our approach splits the usual PAC-Bayesian analysis into two components: (i) choose F to use in Eq. (1), and (ii) obtain an upper bound on the exponential moment $\xi(Q^0)$. The cost of generality is that for each specific choice of the bound (technically, a choice of a function F and Q^0) we need to study the behaviour of the exponential moment $\xi(Q^0)$, and in particular, provide a reasonable, possibly data-dependent upper bound on it. We stress that the only technical step necessary for the introduction of a data-dependent prior is a bound on $\xi(Q^0)$, the rest is taken care of by Eq. (1). We are not aware of previous work making the role of the exponential moment[6] explicit in PAC-Bayesian analysis with data-dependent priors.

2.1 A PAC-Bayes bound with a data-dependent Gibbs prior

First we present a novel approach to data-dependent PAC-Bayes priors, which is based on the prior *empirical Gibbs* distribution $Q^0_s(dh) \propto e^{-\gamma \hat{L}(h,s)}\mu(dh)$ for some fixed $\gamma > 0$ and base measure μ over $\mathcal{H}$. In this approach, to upper-bound the exponential moment, we focus on the specific choice $F(x,y) = \sqrt{n}(x-y)$, and we prove that in this case

$$\log(\xi(Q^0)) \le 2\left(1 + \frac{2\gamma}{\sqrt{n}}\right) + \log\left(1+\sqrt{e}\right) .$$

The proof (Appendix E) is based on the algorithmic stability argument for Gibbs densities, inspired by the proof of [Kuzborskij et al., 2019, Theorem 1]. Combining this with Eq. (1), for any posterior $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over size-n i.i.d. samples S we have

$$Q_S[L] - Q_S[\hat{L}_S] \le \frac{1}{\sqrt{n}}\left(\mathrm{KL}(Q_S\|Q^0_S) + 2\Big(1 + \frac{2\gamma}{\sqrt{n}}\Big) + \log\Big(\frac{1+\sqrt{e}}{\delta}\Big)\right) . \tag{3}$$

Interestingly, the choice $Q = Q^0$ gives the smallest right-hand side in Eq. (3) (however, it does not necessarily minimize the bound on $Q_S[L]$) which leads to the following high-probability bound for the Gibbs learner: $Q_S[L] - Q_S[\hat{L}_S] \lesssim 1/\sqrt{n} + \gamma/n$. Notice that this bound gains an additive $1/\sqrt{n}$ compared to the bound in expectation of Raginsky et al. [2017].

2.2 PAC-Bayes bounds with d-stable data-dependent priors

Next we discuss an approach to convert any PAC-Bayes bound with a usual 'data-free' prior into a bound with a stable data-dependent prior, which is accomplished by generalizing a technique from Dziugaite and Roy [2018b]. In particular, we show (Appendix C) that for any fixed 'data-free' distribution $Q^* \in \mathcal{M}_1(\mathcal{H})$ and stochastic kernel $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{H})$ satisfying the $\mathrm{DP}(\epsilon)$ property[7],

$$\xi(Q^0) \le 2\max\{\xi(Q^*),1\}\exp\left\{\frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\log\left(\frac{2}{\beta}\right)}\right\} \qquad \beta \in (0,1) . \tag{4}$$

Eq. (4) suggests that one should take infimum over 'data-free' distributions Q^* to get the tightest possible bound (and make the bound free from Q^*). Note that different choices of F would lead

[5] Germain et al. [2009] presented a similar generic PAC-Bayes inequality but with fixed 'data-free' priors.

[6] Audibert and Bousquet [2007] separately analyzed the exponential moment but under 'data-free' priors.

[7] See Appendix C.

to different forms of $\xi(Q^*)$ —essentially, upper bounds on the exponential moment typically considered in the PAC-Bayesian literature. For example, taking $F(x,y) = n\,\mathrm{kl}(x\|y)$ one can show that $\xi(Q^*) \le 2\sqrt{n}$ [Maurer, 2004], and by fixing $\beta = 2/3$ we derive a bound that is equivalent to Theorem 4.2 of Dziugaite and Roy [2018b] but with slightly improved constants:

$$\mathrm{kl}(Q_S[\hat{L}_S]\|Q_S[L]) \le \frac{1}{n}\left(\mathrm{KL}(Q_S\|Q_S^0) + \tfrac{1}{2}n\epsilon^2 + \epsilon\sqrt{\tfrac{\log(3)}{2}\,n} + \log(\tfrac{3\sqrt{n}}{\delta})\right) .$$

A more general version of Eq. (4), whose derivation is based on the notion of *max-information* [Dwork et al., 2015a], is discussed in Appendix C (see Lemma 3 there) and proved in Appendix D. The details of the conversion recipe are also in Appendix C.

2.3 Towards a PAC-Bayes bound with a free range loss function

Consider the case that the loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0,\infty)$ has unbounded range. For any $\lambda > 0$ we may upper-bound $\mathbb{E}[\exp\{-\lambda n \hat{L}_n(h,S)\}]$ by standard techniques under the i.i.d. data-generation model. Then with a few calculations we obtain:

$$\mathbb{E}[e^{\lambda n(L(h)-\hat{L}_n(h,s))}] \le e^{\frac{\lambda^2 n}{2}\,\mathbb{E}[\ell(h,Z)^2]} .$$

Then assuming $\infty > M = \sup_h \mathbb{E}[\ell(h,Z)^2]$ (see Holland [2019] whose main result required this), using the function $f(h,s) = \lambda n\big(L(h) - \hat{L}_n(h,s)\big) - \frac{\lambda^2 n}{2}M$ with a usual 'data-free' prior Q^0, the exponential moment satisfies $\xi \le 1$. Thus, for any posterior $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over size-n i.i.d. samples S we have

$$Q_S[L] \le Q_S[\hat{L}_S] + \frac{\mathrm{KL}(Q_S\|Q^0) + \log(1/\delta)}{n\lambda} + \frac{\lambda}{2}\sup_h \mathbb{E}[\ell(h,Z)^2] . \tag{5}$$

This illustrates that PAC-Bayes bounds are possible with unbounded loss functions. However, since the bound given in Eq. (5) is dominated by $M = \sup_h \mathbb{E}[\ell(h,Z)^2]$, this specific example is not the best upper bound that we could hope for. It is the focus of ongoing research to obtain tighter bounds for general hypothesis classes under loss functions with unbounded range.

3 Discussion

The resurgence of the PAC-Bayes approach has been in part motivated by the interest in generalization properties of neural networks. While Langford and Caruana [2001] used a PAC-Bayesian bound to evaluate the error of a (stochastic) neural network classifier, Dziugaite and Roy [2017] obtained numerically non-vacuous generalization bounds by turning a PAC-Bayes bound on the error into a training objective. Several subsequent studies [Blundell et al., 2015, Rivasplata et al., 2019, Mhammedi et al., 2019] took this approach further, sometimes with links to the generalization ability of stochastic optimization [London, 2017, Neyshabur et al., 2018, Dziugaite and Roy, 2018a].

Our work mainly contributes in the direction of connecting PAC-Bayes priors to data. We point out the benefit of separating the proof of the general PAC-Bayes inequality from techniques to bound the exponential moment of the function used in the inequality. This made it possible to derive a PAC-Bayes bound where the prior is data-dependent by default. Obtaining more cases of data-dependent priors is the topic of ongoing research. Our work briefly touched upon boundedness of the loss function, which generally is difficult to avoid in PAC-Bayesian analysis due to the need to control higher moments. While the specific PAC-Bayes bound for loss functions with unbounded range presented here is a rather restricted case, deriving realistic cases is the topic of ongoing research.

Notice that a line of work related to connecting priors to data was explored by Lever et al. [2013], Pentina and Lampert [2014] and more recently by Rivasplata et al. [2018], who assumed that priors are *distribution-dependent*. In that setting priors are still 'data-free' but in a less agnostic fashion (compared to an arbitrary fixed prior), which allows to demonstrate improvements for "nice" data-generating distributions. Finally, it is worth mentioning that the PAC-Bayesian analysis extends beyond bounds on the gap between population and empirical losses: A large body of literature has also looked into upper and lower bounds on the *excess risk*, namely, $Q_S[L] - \inf_{h\in\mathcal{H}} L(h)$, e.g. Catoni [2007], Alquier et al. [2016], Grünwald and Mehta [2019], Kuzborskij et al. [2019]. The approach of analyzing the gap is generally complementary to such excess risk analyses, while on the other hand our results point out interesting directions for future research.

References

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

J.-Y. Audibert and O. Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8(Apr):863–889, 2007.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learing (ICML)*, pages 1613–1622, 2015.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

O. Catoni. PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, 2007. ISSN 07492170. URL http://www.jstor.org/stable/20461499.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in Adaptive Data Analysis and Holdout Reuse. arXiv:1506.02629, 2015a.

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, pages 117–126. ACM, 2015b.

G. K. Dziugaite and D. Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. In *International Conference on Machine Learing (ICML)*, pages 1376–1385, 2018a.

G. K. Dziugaite and D. M. Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

G. K. Dziugaite and D. M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8430–8441, 2018b.

Y. Freund. Self bounding learning algorithms. In *Proceedings of the eleventh annual conference on Computational Learning Theory*, pages 247–258. ACM, 1998.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learing (ICML)*, pages 353–360. ACM, 2009.

P. D. Grünwald and N. A. Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory (ALT)*, volume 98 of *Proceedings of Machine Learning Research*, pages 433–465, Chicago, Illinois, 22–24 Mar 2019. PMLR.

M. J. Holland. PAC-Bayes under potentially heavy tails. In *Conference on Neural Information Processing Systems (NIPS)*, 2019. To appear.

O. Kallenberg. *Random Measures, Theory and Applications*. Springer, 2017.

I. Kuzborskij, N. Cesa-Bianchi, and C. Szepesvári. Distribution-Dependent Analysis of Gibbs-ERM Principle. In A. Beygelzimer and D. Hsu, editors, *Conference on Computational Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 2028–2054, Phoenix, USA, 25–28 Jun 2019. PMLR.

J. Langford. Tutorial on Practical Prediction Theory for Classification. *Journal of Machine Learning Research*, 6(Mar):273–306, 2005.

J. Langford and R. Caruana. (Not) bounding the true error. In *Conference on Neural Information Processing Systems (NIPS)*, pages 809–816, 2001.

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

B. London. A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent. In *Conference on Neural Information Processing Systems (NIPS)*, pages 2931–2940, 2017.

A. Maurer. A note on the PAC Bayesian theorem. arXiv:cs/0411099, 2004.

D. A. McAllester. Some PAC-Bayesian theorems. In *Conference on Computational Learning Theory (COLT)*, pages 230–234. ACM, 1998. Also one year later in *Machine Learning* 37(3), pages 355–363, 1999.

F. McSherry and K. Talwar. Mechanism Design via Differential Privacy. In *FOCS*, volume 7, pages 94–103, 2007.

Z. Mhammedi, P. D. Grunwald, and B. Guedj. PAC-Bayes Un-Expected Bernstein Inequality. arXiv:1905.13367, 2019.

B. Neyshabur, S. Bhojanapalli, and N. Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

A. Pentina and C. H. Lampert. A PAC-Bayesian Bound for Lifelong Learning. In *International Conference on Machine Learing (ICML)*, pages 991–999, 2014.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Conference on Computational Learning Theory (COLT)*, 2017.

O. Rivasplata, E. Parrado-Hernández, J. Shawe-Taylor, S. Sun, and C. Szepesvári. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 9214–9224, 2018.

O. Rivasplata, V. M. Tankasali, and C. Szepesvári. PAC-Bayes with Backprop. arXiv:1908.07380, 2019.

M. Seeger. PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

N. Thiemann, C. Igel, O. Wintenberger, and Y. Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Algorithmic Learning Theory (ALT)*, pages 466–492, 2017.

T. van Erven. PAC-Bayes Mini-tutorial: A Continuous Union Bound. arXiv:1405.1580, 2014.

A Measure-Theoretic Notation

Let $(\mathcal{X}, \Sigma_{\mathcal{X}})$ be a measurable space, i.e. $\mathcal{X}$ is a non-empty set and $\Sigma_{\mathcal{X}}$ is a sigma-algebra of subsets of $\mathcal{X}$. A measure is a countably additive set function $\nu : \Sigma_{\mathcal{X}} \to [0, +\infty]$ such that $\nu(\varnothing) = 0$. We write $\mathcal{M}(\mathcal{X}, \Sigma_{\mathcal{X}})$ for the set of all measures on this space, and $\mathcal{M}_1(\mathcal{X}, \Sigma_{\mathcal{X}})$ for the set of all measures with total mass 1, i.e. probability measures. Actually, when the sigma-algebra where the measure is defined is clear from the context, the notation may be shortened to $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}_1(\mathcal{X})$, respectively. For any measure $\nu \in \mathcal{M}(\mathcal{X})$ and measurable function $f : \mathcal{X} \to \mathbb{R}$, we write $\nu[f]$ to denote the ν-integral of f, so

$$\nu[f] = \int_{\mathcal{X}} f(x)\nu(dx)\,.$$

Thus for instance if X is an $\mathcal{X}$-valued random variable with probability distribution[8] $P \in \mathcal{M}_1(\mathcal{X})$, then $P[f] = \mathbb{E}[f(X)]$ is the expected value.

[8]For sets $A \in \Sigma_{\mathcal{X}}$ the event that the value of X falls within A has probability $\mathbb{P}[X \in A] = P(A)$.

B Our take on the PAC-Bayes inequality

The following results involve hypothesis- and data-dependent functions $f : \mathcal{H} \times \mathcal{S} \to \mathbb{R}$. Notice that the order $\mathcal{H} \times \mathcal{S}$ is immaterial—functions $\mathcal{S} \times \mathcal{H} \to \mathbb{R}$ are treated the same way. If $\rho \in \mathcal{M}_1(\mathcal{H})$ is a 'data-free' distribution, we will write $\rho[f(\cdot, s)]$ to denote the ρ-average of $f(\cdot, s)$ for fixed s, that is, $\rho[f(\cdot, s)] = \int_{\mathcal{H}} f(h, s)\rho(dh)$. When ρ is data-dependent, that is, a stochastic kernel $\rho \in \mathcal{K}(\mathcal{S}, \mathcal{H})$, we may write ρ_s for the distribution over $\mathcal{H}$ corresponding to a fixed s, so $\rho_s(B) = \rho(s, B)$ for $B \in \Sigma_{\mathcal{H}}$, and $\rho_s[f(\cdot, s)] = \int_{\mathcal{H}} f(h, s)\rho_s(dh)$.

The joint distribution over $\mathcal{S} \times \mathcal{H}$ defined by $P \in \mathcal{M}_1(\mathcal{S})$ and $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ is the measure denoted by $P \otimes Q$ that acts on functions $\phi : \mathcal{S} \times \mathcal{H} \to \mathbb{R}$ as follows:

$$(P \otimes Q)[\phi] = \int_{\mathcal{S}} P(ds) \int_{\mathcal{H}} Q(s, dh)[\phi(s, h)] \,.$$

Drawing a random pair $(S, H) \sim P \otimes Q$ is equivalent to drawing $S \sim P$ and drawing $H \sim Q_S$. In this case, with $\mathbb{E}$ denoting the expectation under the joint distribution $P \otimes Q$, the previous display takes the form $\mathbb{E}[\phi(S, H)] = \mathbb{E}[\mathbb{E}[\phi(S, H)|S]]$.

Lemma 1 *Fix a probability measure $P \in \mathcal{M}_1(\mathcal{S})$, a stochastic kernel $Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{H})$, and a measurable function $f : \mathcal{S} \times \mathcal{H} \to \mathbb{R}$, and let*

$$\xi = \int_{\mathcal{S}} \int_{\mathcal{H}} e^{f(s,h)} Q^0_s(dh) P(ds) \,.$$

(i) For any $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of a pair $(S, H) \sim P \otimes Q$ we have

$$f(S, H) \leq \log \left(\frac{dQ_S}{dQ^0_S}(H) \right) + \log(\xi/\delta) \,.$$

(ii) For any $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of $S \sim P$ we have

$$Q_S[f(S, \cdot)] \leq \mathrm{KL}(Q_S \| Q^0_S) + \log(\xi/\delta) \,.$$

This lemma concerns data-dependent distributions over the hypothesis space. Typically Q is called a 'posterior' distribution, and Q^0 is called a 'prior' distribution. Notice that Q^0 is allowed to be data-dependent by default in our approach. To the best of our knowledge, this lemma is new. A key step of the proof involves a change of measure that can be traced back to Csiszár [1975] and Donsker and Varadhan [1975].

Proof Recall that when Y is a positive random variable, by Markov inequality, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have:

$$\log Y \leq \log \mathbb{E}[Y] + \log(1/\delta) \,. \qquad (\star)$$

Let $Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{H})$, and let $\mathbb{E}^0$ denote expectation under the joint distribution $P \otimes Q^0$. Thus if $S \sim P$ and $H \sim Q^0_S$ we then have $\xi = \mathbb{E}^0[\mathbb{E}^0[e^{f(S,H)}|S]]$.

Let $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ and denote by $\mathbb{E}$ the expectation under the joint distribution $P \otimes Q$. Then by a change of measure we may re-write $\xi = \mathbb{E}^0[e^{f(S,H)}]$ as $\xi = \mathbb{E}[e^{\tilde{f}(S,H)}] = \mathbb{E}[e^D]$ with

$$D = \tilde{f}(S, H) = f(S, H) - \log \left(\frac{dQ_S}{dQ^0_S}(H) \right) \,.$$

(i) Applying inequality ($\star$) to $Y = e^D$, with probability at least $1 - \delta$ over the random draw of the pair $(S, H) \sim P \otimes Q$ we get $D \leq \log \mathbb{E}[e^D] + \log(1/\delta)$.

(ii) Notice that $\mathbb{E}[D|S] = Q_S[f(S, \cdot)] - \mathrm{KL}(Q_S \| Q^0_S)$. By Jensen inequality we have $\mathbb{E}[D|S] \leq \log \mathbb{E}[e^D|S]$, while from ($\star$) applied to $Y = \mathbb{E}[e^D|S]$, with probability at least $1 - \delta$ over the random draw of $S \sim P$ we have $\log \mathbb{E}[e^D|S] \leq \log \mathbb{E}[e^D] + \log(1/\delta)$. ■

Notice that Lemma 1 does not restrict the prior to be a fixed 'data-free' distribution. Indeed, Q^0 may be data-dependent by default. Also, the function f is not restricted to have a bounded range.

Suppose the function f is of the form $f = F \circ A$ with $A : \mathcal{S} \times \mathcal{H} \to \mathbb{R}^k$ and $F : \mathbb{R}^k \to \mathbb{R}$ convex. In this case, by Jensen inequality we have $F(Q_s[A(s,\cdot)]) \leq Q_s[F(A(s,\cdot))]$ and Lemma 1(ii) gives:

Theorem 2 *For any $P \in \mathcal{M}_1(\mathcal{S})$, for any $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{H})$, for any positive integer k, for any measurable function $A : \mathcal{S} \times \mathcal{H} \to \mathbb{R}^k$ and convex function $F : \mathbb{R}^k \to \mathbb{R}$, let $f = F \circ A$ and let $\xi = (P \otimes Q^0)[e^f]$ as in Lemma 1. Then for any $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of $S \sim P$ we have*

$$F(Q_S[A(S,\cdot)]) \leq \mathrm{KL}(Q_S \| Q_S^0) + \log(\xi/\delta)\,. \tag{6}$$

In fact, Theorem 2 is valid with any normed space instead of $\mathbb{R}^k$. This result extends the typically used case where $k = 2$ and $A = (\hat{L}(h,s), L(h))$ is the pair consisting of empirical loss and true population loss. Notice also that ξ is the exponential moment (moment generating function at 1) of the function f under the joint distribution $P \otimes Q^0$. Writing $\mathbb{E}^0$ for the expectation under $P \otimes Q^0$, we have $\xi = \mathbb{E}^0[e^{f(S,H)}]$ with randomly drawn $S \sim P$ and $H \sim Q_S^0$.

In contrast to the existing literature on PAC-Bayes bounds, in our Theorem 2 the distribution Q^0 is allowed to be data-dependent by default. Note that a fixed 'data-free' distribution is equivalent to a constant kernel: $Q_s^0 = Q_{s'}^0$ for all $s, s' \in \mathcal{S}$, hence the usual cases are encompassed. The requirement that Q^0 does not depend on data, as in the literature, plays a role when controlling the exponential moment ξ. This is because with a data-free Q^0 we may swap the order of integration:

$$\xi = \int_{\mathcal{S}} \int_{\mathcal{H}} e^{f(h,s)} Q^0(dh) P(ds) = \int_{\mathcal{H}} \int_{\mathcal{S}} e^{f(h,s)} P(ds) Q^0(dh) =: \xi_{\mathrm{swap}}\,.$$

Then bounding ξ proceeds by calculating or bounding ξ_{swap} for which there are readily available techniques (see e.g. Maurer [2004], Germain et al. [2009], van Erven [2014]). Another important aspect of Theorem 2 is the possibility of using losses with unbounded range. Once again, the usual assumption of previous works that losses are bounded (typically with range $[0,1]$) played a role when calculating the ξ term, but as long as it is possible to bound the exponential moment, the restriction of bounded loss function can be removed. This observation may have implications for analysis of learning algorithms e.g. under the square loss or the cross-entropy loss, which are unbounded.

While the derivation of the results in this section follow steps that are well known, as previous works focused on the case of fixed 'data-free' priors, we think that we are the first ones to explicitly point out that the basic PAC-Bayes argument works even with data-dependent priors provided that the exponential moment ξ can be controlled. Thus, developing PAC-Bayes bounds with data-dependent priors is reduced to controlling ξ. We think that this new argument not only leads to a cleaner presentation of existing results, but it also gives rise to improvements on previous results and some new results, as we have demonstrated.

C d-stable data-dependent priors

Let $\pi \in \mathcal{K}(\mathcal{S},\mathcal{H})$ be a stochastic kernel. Recall that $\mathcal{S} = \mathcal{Z}^n$ is the space of size-n samples. When we say that π satisfies the DP property with $\epsilon > 0$ (written $\mathrm{DP}(\epsilon)$ for short) we mean that whenever s and s' differ only at one element, the corresponding distributions over $\mathcal{H}$ satisfy:

$$\frac{d\pi_s}{d\pi_{s'}} \leq e^{\epsilon}\,.$$

This definition goes back to the literature on privacy-preserving methods for data analysis [Dwork et al., 2015b], however, here we are interested in the technical properties only. This condition on the Radon-Nikodym derivative is equivalent to the condition that for all sets $A \in \Sigma_{\mathcal{H}}$, the ratio $\pi(s,A)/\pi(s',A)$ is upper bounded by e^{ϵ}. Thus, the property entails stability of the data-dependent distribution π_s with respect to small changes in the composition of the n-tuple s, hence it is a kind of distributional stability, or d-stability for short.

As noted before, the main challenge in obtaining PAC-Bayes bounds is in controlling the exponential moment $\xi(n)$. In the following we rely on a notion of β-approximate *max-information* [Dwork et al., 2015a,b], which in our context is defined as

$$I_\infty^\beta(S;Q_S^0) = \log \sup_E \frac{\mathbb{P}((S,Q_S^0)\in E)}{\mathbb{P}((S',Q_S^0)\in E)+\beta} \qquad \beta > 0$$

for S, S' independent copies of each other (same distribution). The next lemma, whose proof is in Appendix D, generalizes an idea we learned from Dziugaite and Roy [2018b]:

Lemma 3 (max-information lemma) *Fix $f : \mathcal{S}\times\mathcal{H}\to\mathbb{R}$, $n\in\mathbb{N}$ and $P_n \in \mathcal{M}_1(\mathcal{S})$. Let $\xi_{\mathrm{bd}}(n) = \inf_{Q'\in\mathcal{M}_1(\mathcal{H})}\int\int e^{f(s,h)}Q'(dh)P_n(ds)$. Then for any $Q^0\in\mathcal{K}(\mathcal{S},\mathcal{H})$ and for any $\beta\in(0,1)$ the following bound on $\xi(n)=\int\int e^{f(s,h)}Q_s^0(dh)P_n(ds)$ holds:*

$$\xi(n) \le 2\max\{\xi_{\mathrm{bd}}(n),1\}\exp\left\{I_\infty^\beta(S;Q_S^0)\right\}.$$

The max-information lemma leads to a general recipe for converting a PAC-Bayes bound with a fixed 'data-free' prior into a PAC-Bayes bound with a data-dependent prior. Suppose that for the usual case that $Q^0\in\mathcal{M}_1(\mathcal{H})$ is a fixed 'data-free' prior, for any $Q\in\mathcal{K}(\mathcal{S},\mathcal{H})$ and $\delta\in(0,1)$, with probability $\ge 1-\delta$ over size-n samples $S\sim P_n$, we have

$$F(Q_S[A(S,\cdot)]) \le \mathrm{KL}(Q_S\|Q^0) + \log(\xi_{\mathrm{bd}}(n)/\delta). \tag{7}$$

This is written in the generic framework of Theorem 2 where $f(s,h)=F(A(s,h))$, and $\xi_{\mathrm{bd}}(n)$ is an upper bound on $\xi(n)=\mathbb{E}^0[e^{f(S,H)}]$ valid when Q^0 is a data-free distribution. Then by Lemma 3, for any $Q^0, Q\in\mathcal{K}(\mathcal{S},\mathcal{H})$, for any $\delta\in(0,1)$, with probability $\ge 1-\delta$ over size-n samples $S\sim P_n$, we have

$$F(Q_S[A(S,\cdot)]) \le \mathrm{KL}(Q_S\|Q_S^0) + \log(2\max\{\xi_{\mathrm{bd}}(n),1\}/\delta) + I_\infty^\beta(S;Q_S^0). \tag{8}$$

The following upper bound (see Dwork et al. [2015a, Theorem 20]) on the max-information $I_\infty^\beta(S;Q_S^0)$ is available when the data-dependent Q^0 satisfies $\mathrm{DP}(\epsilon)$:

$$I_\infty^\beta(S;Q_S^0) \le \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\log(\frac{2}{\beta})}.$$

Therefore, via the max-information lemma, one may derive PAC-Bayes bounds which are valid for d-stable data-dependent priors. More specialized forms of the upper bound can be obtained when a specific form of $\xi_{\mathrm{bd}}(n)$ is available. For instance, starting from the PAC-Bayes-kl bound (Seeger [2002], see also Langford [2005]) we derive the following:

Theorem 4 *For any n, for any $P_1\in\mathcal{M}_1(\mathcal{Z})$, for any $Q^0\in\mathcal{K}(\mathcal{S},\mathcal{H})$ satisfying $\mathrm{DP}(\epsilon)$, for any loss function with range $[0,1]$, for any $Q\in\mathcal{K}(\mathcal{S},\mathcal{H})$, for any $\delta\in(0,1)$, with probability $\ge 1-\delta$ over size-n i.i.d. samples $S\sim P_1^n$ we have*

$$\mathrm{kl}(Q_S[\hat{L}_S]\|Q_S[L]) \le \frac{\mathrm{KL}(Q_S\|Q_S^0) + \log(\frac{3\sqrt{n}}{\delta}) + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\log(3)}}{n}. \tag{9}$$

This is essentially equivalent to [Dziugaite and Roy, 2018b, Theorem 4.2] but with slightly improved constants. The proof of Theorem 4 is as follows.

Under the restrictions of the theorem, we may use $\xi_{\mathrm{bd}}(n)=2\sqrt{n}$ (as per Maurer [2004]) when the prior is a fixed 'data-free' distribution. Then by Lemma 3 we get $\xi(n)\le 2\sqrt{n}e^{I_\infty^\beta(S;Q_S^0)}+\beta$ when the prior is data-dependent. Thus $\xi(n)\le 3\sqrt{n}e^{I_\infty^\beta(S;Q_S^0)}$, which gives

$$\log(\xi(n)) \le \log(3\sqrt{n}) + I_\infty^\beta(S;Q_S^0).$$

On the other hand, as mentioned above, if Q^0 satisfies the $\mathrm{DP}(\epsilon)$ property, then for any $\beta\in(0,1)$ we have the upper bound

$$I_\infty^\beta(S;Q_S^0) \le \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\log(\frac{2}{\beta})}.$$

This is Dwork et al. [2015a, Theorem 20]. Using $\beta=2/3$ completes the proof.

D Proof of the max-information lemma

Let $f(s,h)$ be a data-dependent and hypothesis-dependent function. Recall that s summarizes a size-n sample. Suppose $\xi_{\mathrm{bd}}(n)$ is an upper bound on $\xi(n) = \mathbb{E}^0[e^{f(S,H)}]$ which is valid when $Q^0 \in \mathcal{M}_1(\mathcal{H})$ is fixed (not data-dependent). Now suppose $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{H})$ is a stochastic kernel, so each random size-n data set S maps to a data-dependent distribution Q^0_S over $\mathcal{H}$. The corresponding β-approximate max-information as defined by Dwork et al. [2015a] (see also Dwork et al. [2015b]) is denoted $I^\beta_\infty(S;Q^0_S)$ in our context. The max-information argument to bound $\xi(n)$ goes as follows:

$$\begin{aligned}
\xi(n) &= \int_{\mathcal{S}}\int_{\mathcal{H}} e^{f(s,h)} Q^0_s(dh) P_n(ds) \\
&\leq e^{I^\beta_\infty(S;Q^0_S)} \int_{\mathcal{S}}\int_{\mathcal{S}}\int_{\mathcal{H}} e^{f(s,h)} Q^0_{s'}(dh) P_n(ds) P_n(ds') + \beta \\
&\leq e^{I^\beta_\infty(S;Q^0_S)} \xi_{\mathrm{bd}}(n) + \beta\,.
\end{aligned}$$

The first inequality, valid for any $\beta \in (0,1)$, is due to the definition of $I^\beta_\infty(S;Q^0_S)$. The second inequality is due to the fact that $f(s,h)$ and $Q^0_{s'}$ have been decoupled, so that for each fixed $s' \in \mathcal{S}$ the internal double integral is upper bounded by $\xi_{\mathrm{bd}}(n)$.

Thus we get $\xi(n) \leq 2\max\{\xi_{\mathrm{bd}}(n), 1\} e^{I^\beta_\infty(S;Q^0_S)}$ by considering the cases $\xi_{\mathrm{bd}}(n) \leq 1$ and $\xi_{\mathrm{bd}}(n) > 1$. This finishes the proof of the "max-information lemma" (Lemma 3).

Notice that if a data-dependent prior $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{H})$ satisfies $\mathrm{DP}(\epsilon)$ for some $\epsilon > 0$, then in the exponential moment

$$\xi(n) = \int_{\mathcal{S}}\int_{\mathcal{H}} e^{f(h,s)} Q^0_s(dh) P_n(ds)$$

we may change the measure Q^0_s to $Q^0_{s'}$ with any fixed $s' \in \mathcal{S}$, and the Radon-Nikodym derivative satisfies $dQ^0_s / dQ^0_{s'} \leq e^{n\epsilon}$, so we have

$$\xi(n) \leq e^{n\epsilon} \int_{\mathcal{S}}\int_{\mathcal{H}} e^{f(h,s)} Q^0_{s'}(dh) P_n(ds) \leq e^{n\epsilon} \xi_{\mathrm{bd}}(n)$$

where the integral on the right hand side is upper bounded by $\xi_{\mathrm{bd}}(n)$ since $Q^0_{s'}$ is now a fixed distribution (constant kernel). Thus the max-information lemma gives a refined analysis leading to an upper bound on $\xi(n)$ where '$n\epsilon$' is replaced with $I^\beta_\infty(S;Q^0_S)$.

E Proof of the bound for data-dependent Gibbs priors

For the sake of clarity let us recall once more that $P \otimes Q$ denotes the joint distribution over $\mathcal{S}\times\mathcal{H}$ defined by $P \in \mathcal{M}_1(\mathcal{S})$ and $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$. Drawing a random pair $(S,H) \sim P \otimes Q$ is equivalent to drawing $S \sim P$ and drawing $H \sim Q_S$. With $\mathbb{E}$ denoting expectation under $P \otimes Q$, for measurable functions $\phi : \mathcal{S}\times\mathcal{H} \to \mathbb{R}$ we have $\mathbb{E}[\phi(S,H)] = \mathbb{E}[\mathbb{E}[\phi(S,H)|S]]$.

Lemma 5 *For any n, for any loss function with range $[0,b]$, for any $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ such that $Q_s(dh) \propto e^{-\gamma \hat{L}_n(h,s)}\mu(dh)$, the following upper bound on $\xi(n) = \mathbb{E}[e^{\sqrt{n}\left(L(H)-\hat{L}_n(H,S)\right)}]$ holds:*

$$\log(\xi(n)) \leq 2b^2\left(1 + \frac{2\gamma}{\sqrt{n}}\right) + \log\left(1 + e^{b^2/2}\right)\,.$$

For the proof of Lemma 5, we will use the shorthand $\Delta_s(h) = \sqrt{n}\big(L(h) - \hat{L}_n(h,s)\big)$ where $(s,h) \in \mathcal{S}\times\mathcal{H}$. We need two technical results, quoted next for convenience.

Lemma 6 (Boucheron et al. 2013, Lemma 4.18) *Let S be a real-valued integrable random variable such that*

$$\log \mathbb{E}\left[e^{\alpha(S-\mathbb{E}[S])}\right] \leq \frac{\alpha^2\sigma^2}{2} \qquad \alpha > 0$$

holds for some $\sigma > 0$, and let S' be another real-valued integrable random variable. Then we have $\mathbb{E}[S'] - \mathbb{E}[S] \leq \sqrt{2\sigma^2\,\mathrm{KL}\left(\mathrm{Law}(S') \| \mathrm{Law}(S)\right)}$.

Lemma 7 (Kuzborskij et al. 2019, Lemma 9) *Let $f_A, f_B : \mathcal{H} \to \mathbb{R}$ be measurable functions such that the normalizing factors*

$$N_A = \int_{\mathcal{H}} e^{-\gamma f_A(h)} \, \mathrm{d}h \qquad \text{and} \qquad N_B = \int_{\mathcal{H}} e^{-\gamma f_B(h)} \, \mathrm{d}h$$

are finite for all $\gamma > 0$, and let p_A and p_B be the corresponding densities:

$$p_A(h) = \frac{1}{N_A} e^{-\gamma f_A(h)} , \qquad p_B(h) = \frac{1}{N_B} e^{-\gamma f_B(h)} , \qquad h \in \mathcal{H} .$$

Whenever $N_A > 0$ we have that

$$\ln\left(\frac{N_B}{N_A}\right) \leq \gamma \int_{\mathcal{H}} p_B(h) \left(f_A(h) - f_B(h)\right) \mathrm{d}h .$$

The last lemma is helpful for bounding the log-ratio of Gibbs integrals. The notation '$\mathrm{d}h$' stands for integration with respect to a fixed reference measure (suppressed in the notation) over the space $\mathcal{H}$. Now we are ready for the proof.

Proof [of Lemma 5] Throughout the proof we will use an auxiliary random variable H' drawn randomly from a distribution $Q' \in \mathcal{M}_1(\mathcal{H})$ that does not depend on S in any way. The first step is to relate the exponential moment of $\Delta_S(H)$ to the expectation of $\Delta_S(H)$ under a suitably defined Gibbs distribution and the exponential moment of $\Delta_S(H')$. Then the expectation of $\Delta_S(H)$ will be bounded via an *algorithmic stability* analysis of the Gibbs density as in the proof of Theorem 1 by Kuzborskij et al. [2019], while the exponential moment of $\Delta_S(H')$ is bounded by readily available techniques since the distribution of H' is decoupled from S.

We will carry out the first step through the continuous version of the log-sum inequality, which says that for positive random variables A and B one has:

$$\mathbb{E}[A] \ln \frac{\mathbb{E}[A]}{\mathbb{E}[B]} \leq \mathbb{E}\left[A \ln\left(\frac{A}{B}\right)\right] .$$

We will use this inequality with the random variables $A = e^{\Delta_S(H)}$ and $B = e^{(\Delta_S(H'))_+}$ where $(x)_+ = x\mathbf{1}_{x \geq 0}$ is the positive part function. This gives

$$\mathbb{E}\left[e^{\Delta_S(H)}\right] \left(\ln \mathbb{E}\left[e^{\Delta_S(H)}\right] - \ln \mathbb{E}\left[e^{(\Delta_S(H'))_+}\right]\right) \leq \mathbb{E}\left[e^{\Delta_S(H)} \left(\Delta_S(H) - (\Delta_S(H'))_+\right)\right]$$

so then rearranging

$$\begin{aligned} \ln \mathbb{E}\left[e^{\Delta_S(H)}\right] &\leq \mathbb{E}\left[\frac{e^{\Delta_S(H)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]} \left(\Delta_S(H) - (\Delta_S(H'))_+\right)\right] + \ln \mathbb{E}\left[e^{(\Delta_S(H'))_+}\right] \\ &\leq \mathbb{E}\left[\frac{e^{\Delta_S(H)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]} \Delta_S(H)\right] + \ln \mathbb{E}\left[e^{(\Delta_S(H'))_+}\right] . \end{aligned} \tag{10}$$

Let's write q_s for the density of Q_s with respect to a reference measure $\mathrm{d}h$ over $\mathcal{H}$, and introduce a measure

$$\mathrm{d}\mu_S(h) = \frac{e^{\Delta_S(h)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]} \, \mathrm{d}q_S(h) \quad h \in \mathcal{H} .$$

Then the inequality (10) can be written as

$$\ln \mathbb{E}\left[e^{\Delta_S(H)}\right] \leq \underbrace{\mathbb{E} \int \Delta_S(h) \, \mathrm{d}\mu_S(h)}_{(I)} + \underbrace{\ln \mathbb{E}\left[e^{(\Delta_S(H'))_+}\right]}_{(II)} .$$

Bounding (I). We handle the first term through the stability analysis of the density μ_S. By $S^{(i)} = (Z_{1:i-1}, Z'_1, Z_{i+1:n})$ we denote the sample obtained from $S = (Z_{1:i-1}, Z_i, Z_{i+1:n})$ by replacing the

ith entry with an independent copy Z_1'. In particular,

$$\begin{aligned}\frac{1}{\sqrt{n}}\,\mathbb{E}\int \Delta_S(h)\,\mathrm{d}\mu_S(h) &= \mathbb{E}\int \ell(h,Z_1')\,\mathrm{d}\mu_S(h) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\int \ell(h,Z_i)\,\mathrm{d}\mu_S(h)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\int \left(\ell(h,Z_1')-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h) \qquad (11)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\int \ell(h,Z_i)\,\mathrm{d}\mu_{S^{(i)}}(h) - \int \ell(h,Z_i)\,\mathrm{d}\mu_S(h)\right].\end{aligned}$$

The last equality comes from switching Z_1' and Z_i since these variables are distributed identically. Now we use Lemma 6 with $\mu_{S^{(i)}}$ and μ_S, and with $\sigma = b$, to get that

$$\int \ell(h,Z_i)\,\mathrm{d}\mu_{S^{(i)}}(h) - \int \ell(h,Z_i)\,\mathrm{d}\mu_S(h) \le \sqrt{2b^2\,\mathrm{KL}\left(\mu_{S^{(i)}}\|\mu_S\right)}\,.$$

Notice that we may use $\sigma = b$ in Lemma 6 since the loss function has range $[0,b]$. Focusing on the KL-divergence, and writing '$\mathrm{d}h$' for a reference measure on $\mathcal{H}$ with respect to which q_S, μ_S, $\mu_{S^{(i)}}$ are absolutely continuous,

$$\begin{aligned}\mathrm{KL}\left(\mu_{S^{(i)}}\|\mu_S\right) &= \int \ln(\mathrm{d}\mu_{S^{(i)}}(h)/\,\mathrm{d}h)\,\mathrm{d}\mu_{S^{(i)}}(h) - \int \ln(\mathrm{d}\mu_S(h)/\,\mathrm{d}h)\,\mathrm{d}\mu_{S^{(i)}}(h)\\ &= \int \ln\left(\frac{e^{\Delta_{S^{(i)}}(h)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]}\frac{e^{-\gamma\hat{L}_{S^{(i)}}(h)}}{N_{S^{(i)}}}\right)\mathrm{d}\mu_{S^{(i)}}(h) - \int \ln\left(\frac{e^{\Delta_S(h)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]}\frac{e^{-\gamma\hat{L}_S(h)}}{N_S}\right)\mathrm{d}\mu_{S^{(i)}}(h)\\ &= \int \left(\Delta_{S^{(i)}}(h)-\Delta_S(h)\right)\mathrm{d}\mu_{S^{(i)}}(h) + \ln\left(\frac{N_S}{N_{S^{(i)}}}\right) + \gamma\int\left(\hat{L}_S(h)-\hat{L}_{S^{(i)}}(h)\right)\mathrm{d}\mu_{S^{(i)}}(h)\\ &\le \sqrt{n}\int\left(\hat{L}_S(h)-\hat{L}_{S^{(i)}}(h)\right)\mathrm{d}\mu_{S^{(i)}}(h) \qquad \text{(By definition of } \Delta_S)\\ &\quad + \gamma\int\left(\hat{L}_{S^{(i)}}(h)-\hat{L}_S(h)\right)\mathrm{d}\mu_S(h) \qquad \text{(By Lemma 7)}\\ &\quad + \gamma\int\left(\hat{L}_S(h)-\hat{L}_{S^{(i)}}(h)\right)\mathrm{d}\mu_{S^{(i)}}(h)\\ &= \frac{1}{\sqrt{n}}\int\left(\ell(h,Z_i)-\ell(h,Z_1')\right)\mathrm{d}\mu_{S^{(i)}}(h)\\ &\quad + \frac{\gamma}{n}\int\left(\ell(h,Z_1')-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h)\\ &\quad + \frac{\gamma}{n}\int\left(\ell(h,Z_i)-\ell(h,Z_1')\right)\mathrm{d}\mu_{S^{(i)}}(h)\,,\end{aligned}$$

where the last step is due to multiple cancellations. Therefore, taking expectation,

$$\mathbb{E}\,\mathrm{KL}\left(\mu_{S^{(i)}}\|\mu_S\right) \le \left(\frac{1}{\sqrt{n}}+\frac{2\gamma}{n}\right)\mathbb{E}\left[\int\left(\ell(h,Z_1')-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h)\right].$$

Putting all together, for each term in Eq. (11) (each $i\in[n]$) we get

$$\begin{aligned}&\mathbb{E}\left[\int\left(\ell(h,Z_1')-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h)\right] = \mathbb{E}\left[\int\ell(h,Z_i)\,\mathrm{d}\mu_{S^{(i)}}(h) - \int\ell(h,Z_i)\,\mathrm{d}\mu_S(h)\right]\\ &\le \mathbb{E}\left[\sqrt{2b^2\,\mathrm{KL}\left(\mu_{S^{(i)}}\|\mu_S\right)}\right] \le \sqrt{2b^2\,\mathbb{E}[\mathrm{KL}\left(\mu_{S^{(i)}}\|\mu_S\right)]} \qquad \text{(By Lemma 6 and Jensen)}\\ &= \sqrt{2b^2\left(\frac{1}{\sqrt{n}}+\frac{2\gamma}{n}\right)\mathbb{E}\left[\int\left(\ell(h,Z_1')-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h)\right]}\,.\end{aligned}$$

The last calculation implies

$$\left|\mathbb{E}\left[\int\left(\ell(h,Z_i)-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h)\right]\right| \le 2b^2\left(\frac{1}{\sqrt{n}}+\frac{2\gamma}{n}\right).$$

Finally, combining this with Eq. (11) gives

$$\mathbb{E}\int\Delta_S(h)\,\mathrm{d}\mu_S(h) \le 2b^2\left(1+\frac{2\gamma}{\sqrt{n}}\right). \qquad (12)$$

Bounding (II)**.** Now we turn our attention to the exponential moment of $(\Delta_S(H'))_+$ in (10):

$$\begin{aligned}\ln \mathbb{E}\left[e^{(\Delta_S(H'))_+}\right] &= \ln \mathbb{E}\,\mathbb{E}\left[e^{(\Delta_S(H'))_+} \mid S\right] \\ &= \ln \mathbb{E}\,\mathbb{E}\left[e^{(\Delta_S(H'))_+} \mid H'\right] \qquad \text{(swapping the order of integration)}\end{aligned}$$

and observe that the internal expectation is bounded as

$$\begin{aligned}\mathbb{E}\left[e^{(\Delta_S(H'))_+} \mid H'\right] &\leq 1 + \mathbb{E}\left[e^{\Delta_S(H')} \mid H'\right] \\ &= 1 + \mathbb{E}\left[\exp\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbb{E}[\ell(H', Z_1') \mid H'] - \ell(H', Z_i)\right)\right) \mid H'\right] \\ &= 1 + \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{1}{\sqrt{n}}\left(\mathbb{E}[\ell(H', Z_1') \mid H'] - \ell(H', Z_i)\right)\right) \mid H'\right] \\ &\leq 1 + \prod_{i=1}^{n}\exp\left(\left(2b/\sqrt{n}\right)^2/8\right) \\ &= 1 + e^{b^2/2}\,,\end{aligned}$$

where we obtain the last inequality thanks to the Hoeffding's lemma for independent random variables between $[-b/\sqrt{n}, b/\sqrt{n}]$. Plugging bounds on terms (I) and (II) into Eq. (10) finishes the proof of Lemma 5. ■

We obtain the following generalization bound by observing that the Gibbs distribution with density $\propto e^{-\gamma \hat{L}_n(h,s)}$ satisfies the $\mathrm{DP}(2\gamma/n)$ property.

Corollary 8 *For any n, for any $P_1 \in \mathcal{M}_1(\mathcal{Z})$, for any loss function with range $[0,1]$, for any $\gamma > 0$, for any $Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ such that $Q_s^0 \propto e^{-\gamma \hat{L}_n(h,s)}$, for any $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ and $\delta \in (0,1)$, with probability $\geq 1-\delta$ over size-n i.i.d. samples $S \sim P_1^n$, we have*

$$|Q_S[\hat{L}_S] - Q_S[L]| \leq \sqrt{\frac{\mathrm{KL}(Q_S\|Q_S^0)}{2n}} + \frac{\gamma}{n} + \sqrt[4]{\frac{1}{2}\log(3)}\,\frac{\sqrt{\gamma}}{n^{\frac{3}{4}}} + \sqrt{\frac{\log\left(\frac{3\sqrt{n}}{\delta}\right)}{2n}}\,.$$

Proof Theorem 6 of McSherry and Talwar [2007] gives that the Gibbs distribution $Q_s^0 \propto e^{-\gamma \hat{L}(h,s)}$ with potential satisfying $\sup_{s,s'}\sup_{h\in\mathcal{S}} \hat{L}_s(h) - \hat{L}_{s'}(h) \leq 1/n$ for $s, s' \in \mathcal{S}$ that differ at most in one entry, satisfies $\mathrm{DP}(2\gamma/n)$. Combined with Theorem 4, this gives

$$\mathrm{kl}(Q_S[\hat{L}_S]\|Q_S[L]) \leq \frac{1}{n}\left(\mathrm{KL}(Q_S\|Q_S^0) + \frac{2\gamma^2}{n} + \sqrt{2\log(3)}\frac{\gamma}{\sqrt{n}} + \log\left(\tfrac{3\sqrt{n}}{\delta}\right)\right)$$

and applying Pinsker's inequality $2(p-q)^2 \leq \mathrm{kl}(p\|q)$ and sub-additivity of $t \mapsto \sqrt{t}$:

$$\begin{aligned}|Q_S[\hat{L}_S] - Q_S[L]| &\leq \frac{1}{\sqrt{2n}}\sqrt{\mathrm{KL}(Q_S\|Q_S^0) + \frac{2\gamma^2}{n} + \sqrt{2\log(3)}\frac{\gamma}{\sqrt{n}} + \log\left(\tfrac{3\sqrt{n}}{\delta}\right)} \\ &\leq \sqrt{\frac{\mathrm{KL}(Q_S\|Q_S^0)}{2n}} + \frac{\gamma}{n} + \sqrt[4]{\frac{1}{2}\log(3)}\,\frac{\sqrt{\gamma}}{n^{\frac{3}{4}}} + \sqrt{\frac{\log\left(\frac{3\sqrt{n}}{\delta}\right)}{2n}}\,.\end{aligned}$$

■

While the argument based on d-stability (i.e. Corollary 8) gives a result where the order in γ/n matches the one in our bound for the empirical Gibbs prior, our analysis offers an alternative proof technique that might be of independent interest.