PAC-Bayes Analysis Beyond the Usual Bounds

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 quality of predictions 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 basic PAC-Bayes inequality for stochastic kernels, 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 illustrate the use of data-dependent priors. We also present a simple bound that is valid for a loss function with unbounded range. Our analysis clarifies that those two requirements were used to upper-bound an exponential moment term, while the basic 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 $S = Z^n$ (e.g. $Z = \mathbb{R}^d \times \mathcal{Y}$) according to some unknown probability distribution† $P_n \in \mathcal{M}_1(S)$. Typically $Z_1, \ldots, Z_n$ are independent and share a common distribution $P_1 \in \mathcal{M}_1(Z)$. 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 $Q[L] = \int_\mathcal{H} L(h)Q(\text{d}h)$, i.e. the $Q$-average of the standard population loss $L(h) = \int \ell(h, z)P_1(\text{d}z)$ for a fixed hypothesis $h \in \mathcal{H}$, where $\ell : \mathcal{H} \times \mathcal{Z} \to [0, \infty)$ is a given loss function and $P_1 \in \mathcal{M}_1(Z)$ generates one random example. Similarly, the expected empirical loss is $Q[\hat{L}_s] = \int_\mathcal{H} \hat{L}_s(h)Q(\text{d}h)$, 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 formalizing “data-dependent distributions over $\mathcal{H}$” in a way that makes explicit their difference to fixed “data-free” distributions over $\mathcal{H}$.

†We write $\mathcal{M}_1(X)$ to denote the family of probability measures over a set $X$, see Appendix A.
Randomised predictors with a data-dependent distribution. A data-dependent distribution over the space $\mathcal{H}$ is formalized as a stochastic kernel\(^2\) from $\mathcal{S}$ to $\mathcal{H}$, which is defined as a mapping $Q : \mathcal{S} \times \Sigma_{\mathcal{H}} \to [0, 1]$ 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 stochastic kernels from $\mathcal{S}$ to —distributions over— $\mathcal{H}$. We will reserve the notation $\mathcal{M}_1(\mathcal{H})$ for the set of ‘data-free’ 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 $Q_s[\hat{L}_s] = \int \hat{L}_s(h)Q_s(dh)$ for the expected population loss and the 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 $Q_S(dh) \propto e^{-\gamma L(h, S)}\mu(dh)$ for some $\gamma > 0$, with $\mu$ 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 study the behaviour of a learning method on different distributions— for example, some data-generating distributions might give faster convergence rates than others. Finally, data-dependent bounds are expressed in terms of empirical quantities that can be computed directly from data. These are useful for building and comparing predictors [Catoni, 2007], and also for “self-bounding” learning algorithms, which are algorithms that use all the data to simultaneously provide a predictor and a risk certificate that is valid on unseen examples [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 ‘data-free’ 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 learning, 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 so it may be different from the posterior that would be obtained from $Q^0$ through Bayesian inference. In this sense, the PAC-Bayes approach affords an extra level of flexibility in the choice of distributions, even compared to generalized Bayesian approaches [Bissiri et al., 2016].

The focus of PAC-Bayes analysis is deriving bounds on the gap between $Q_S[L]$ and $Q_S[\hat{L}_S]$. For instance, the classical result of McAllester [1999] says the following: For a fixed ‘data-free’ distribution $Q^0 \in \mathcal{M}_1(\mathcal{H})$, bounded loss function with range $[0, 1]$, stochastic kernel $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ and for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over size-$n$ random samples $S$:

\[
Q_S[L] - Q_S[\hat{L}_S] \leq \sqrt{\frac{1}{2n - 1} \left( \text{KL}(Q_S||Q^0) + \log \left( \frac{\text{size} \cdot 2^n}{\delta} \right) \right)},
\]

where $\text{KL}$ stands for the Kullback-Leibler divergence\(^4\). In words, Eq. (1) tells us that the population loss is controlled by a trade-off between the empirical loss and the deviation of the posterior from the prior as captured by the KL divergence. This result is usually presented under a statement that says that with probability at least $1 - \delta$, the above inequality holds for all probability measures $Q$ over $\mathcal{H}$ with $Q$ replacing $Q_S$. Since that formulation hides the data-dependence, while our main interest is in results for data-dependent distributions (contrasted to results for fixed ‘data-free’ distributions), we argue that the formulation in terms of stochastic kernels is to be preferred.

---

\(^2\)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: $\text{KL}(Q||Q') = \int Q \log \left( \frac{dQ}{dQ'} \right) dQ$, where $dQ/dQ'$ denotes the Radon-Nikodym derivative.
A large body of subsequent work focused on refining the PAC-Bayes analysis by means of alternative proof techniques and different ways to measure the gap between $Q_S[L]$ and $Q_S[L_S]$. For instance, Seeger [2002], improving over Langford and Seeger [2001], derived a PAC-Bayes bound on the binary KL divergence commonly called the PAC-Bayes-kl bound:

$$\text{kl}(Q_S[L_S] \parallel Q_S[L]) \leq \frac{1}{n} \left( \text{KL}(Q_S||Q^0) + \log \left( \frac{n+1}{\delta} \right) \right).$$  

This inequality is tighter than Eq. (1) due to Pinsker’s inequality $2(p - q)^2 \leq \text{kl}(p||q)$. In fact, by a stronger form of Pinsker’s inequality, namely $(p - q)^2 / (2q) \leq \text{kl}(p||q)$ which is valid for $p < q$, from Eq. (2) one easily obtains a localised inequality (see Eq. (6) of McAllester [2003]), which again holds with high probability:

$$Q_S[L] - Q_S[L_S] \lesssim \sqrt{\frac{Q_S[L_S]}{n} \text{KL}(Q_S||Q^0) + \frac{1}{n} \text{KL}(Q_S||Q^0)}.$$

Obviously, the upper bound in Eq. (3) is dominated by the lower-order (second) term whenever the empirical loss $Q_S[L_S]$ is small enough, which makes this inequality very appealing for learning problems based on empirical risk minimization, where the empirical loss is driven to zero. At a high level, such kinds of data-dependent upper bounds on the generalization gap are much desirable, as their empirical terms are closely linked to —and hopefully will help capture more properties of—the data. In this direction valuable contributions were made by Tolstikhin and Seldin [2013] who obtained an empirical PAC-Bayes bound similar in spirit to Eq. (3), but controlled by the sample variance of the loss. An alternative direction to get sharper empirical bounds was explored through tunable bounds (Catoni [2007], van Erven [2014], Thiemann et al. [2017]), which involve a free parameter offering a trade-off between the empirical error term and the KL(Posterior||Prior) term.

Despite their variety and attractive properties, the results discussed above (and the vast majority of the literature) share two crucial limitations: the prior $Q^0$ cannot depend on the training data $S$ and the loss function has to be bounded. It is conceivable that in many realistic situations the population loss is effectively controlled by the KL “complexity” term—indeed, in most modern learning scenarios (e.g. training deep neural networks) the empirical loss is driven to zero. At the same time, the choice of a fixed ‘data-free’ prior essentially becomes a wild guess on how the posterior will look like. Therefore, allowing priors to be data-dependent introduces much needed flexibility, since this opens up the possibility to minimize upper bounds on $Q_S[L]$ in both the posterior and the prior, which should lead to tighter empirical bounds and hence better risk certificates.

These limitations have been removed in the PAC-Bayesian literature in special cases. For instance, Ambroladze et al. [2007] and Parrado-Hernández et al. [2012] used priors which were trained on a held-out portion of the available data, thus enabling empirical bounds by allowing priors to be data-dependent, but independent from the training set. Priors that depend on the full training set have also been studied recently. Thiemann et al. [2017] proposed to construct a prior as a mixture of point masses at a finite number of data-dependent hypotheses trained on a $k$-fold split of the training set, effectively a data-dependent prior. Another approach was proposed by Dziugaite and Roy [2018b]: rather than splitting the training data, they require the data-dependent prior $Q^0_s$ (where $Q^0_s \in \mathcal{K}(S, \mathcal{H})$) to be stable with respect to ‘small’ changes in the composition of the $n$-tuple $s$. As we will see shortly, our main contributions show the benefit of removing these limitations.

In this paper the main focus will be on data-dependent priors in the PAC-Bayes analysis. We will also discuss a particular bound for linear regression with the square loss, to illustrate that the restriction of bounded losses is not mandatory in the PAC-Bayes analysis. Again, we point out that the proof of the basic PAC-Bayes inequality (Theorem 1 below) does not require fixed ‘data-free’ priors or bounded loss functions. Nor does its consequence, Theorem 2, which is a general template for deriving many PAC-Bayes bounds. We will discuss three PAC-Bayes bounds with data-dependent priors, two of which are novel, one of which is for the unbounded square loss function. It is the topic of ongoing research to investigate further learning problems with unbounded losses in the PAC-Bayes analysis, and further examples of data-dependent PAC-Bayes priors.

---

5 The binary KL divergence is defined for $q, q' \in [0, 1]$ as follows: $\text{kl}(q||q') = q \log \left( \frac{q}{q'} \right) + (1 - q) \log \left( \frac{1 - q}{1 - q'} \right)$. In other words, this is the KL between the Bernoulli distributions with parameters $q$ and $q'$.

6 For $x, b, c$ nonnegative, $x \leq c + b \sqrt{x}$ implies $x \leq c + b \sqrt{c} + b^2$.

7 The notation $\lesssim$ hides universal constants and logarithmic factors.
2 Our Contributions

In this paper we discuss a basic PAC-Bayes inequality (Theorem 1 below) and a general template for PAC-Bayesian bounds (Theorem 2 below) encompassing many usual bounds which appear in the literature [McAllester, 1998, Seeger, 2002, Catoni, 2007, Thiemann et al., 2017], but the formulation discussed here, based on representing data-dependent distributions as stochastic kernels, allows the PAC-Bayes priors to be data-dependent by default (see Section 3). Interestingly, the analysis clarifies that it is also possible for the loss functions to have an unbounded range.

Our general PAC-Bayes theorem\(^8\) for stochastic kernels (Theorem 2 in Section 3), in specialized form, implies that for any convex function \( F : \mathbb{R}^2 \to \mathbb{R} \), stochastic kernels \( Q, Q^0 \in \mathcal{K}(S, \mathcal{H}) \) and \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \) over the random draw of data \( S \) one has

\[
F(Q_S[L], Q_S[\hat{L}_S]) \leq KL(Q_S||Q^0_S) + \log(\xi(Q^0)/\delta) ,
\]

where \( \xi(Q^0) \) is the exponential moment of \( F(L(h), \hat{L}_S(h)) \), which is defined as follows:

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

Observe that Eq. (4) is defined for an arbitrary convex function \( F \). This way the usual bounds are encompassed: \( F(x, y) = 2n(x - y)^2 \) yields a McAllester [1998]-type bound, \( F(x, y) = n \operatorname{kl}(y||x) \) gives the bound of Seeger [2002], and \( F(x, y) = n \log\left(\frac{1}{1-x(1-x-y)}\right) - n \lambda y \) gives the bound of Catoni [2007]. Furthermore, \( F(x, y) = n(x - y)^2/(2x) \) leads to the bound of Thiemann et al. [2017], or by a different derivation to the bound of Rivasplata et al. [2019] that holds under the usual requirements of fixed ‘data-free’ distribution \( Q^0 \) and losses within the \([0, 1]\) range:

\[
Q_S[L] \leq \sqrt{Q_S[\hat{L}_S]} + \frac{\text{KL}(Q_S||Q^0_S) + \log\left(\frac{2\sqrt{e}}{\delta}\right)}{2n} + \sqrt{\frac{\text{KL}(Q_S||Q^0) + \log\left(\frac{2\sqrt{e}}{\delta}\right)}{2n}} .
\]

As consequence of the universality of Eq. (4), 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. (4), 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. (4). We are not aware of previous work making the role of the exponential moment explicit\(^9\) in PAC-Bayesian analysis with data-dependent priors.

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

Choosing as prior an empirical Gibbs distribution \( Q^0_S(dh) \propto e^{-\gamma L(h, s)} \mu(dh) \) for some fixed \( \gamma > 0 \) and base measure \( \mu \) over \( \mathcal{H} \), we derive a novel PAC-Bayes bound. The same Gibbs distribution was used by Catoni [2007, Theorem 1.3.1]. Recall that \( s \) is the size-\( n \) sample. We focus on the specific choice \( F(x, y) = \sqrt{n}(x - y) \), and we prove that in this case the exponential moment \( \xi(Q^0) \) satisfies

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

The proof (Appendix B) 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. (4), for any posterior \( Q \in \mathcal{K}(S, \mathcal{H}) \) and \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \) over size-\( n \) i.i.d. samples \( S \) we have

\[
Q_S[L] - Q_S[\hat{L}_S] \leq \frac{1}{\sqrt{n}} \left( \text{KL}(Q_S||Q^0_S) + 2 \left( 1 + \frac{2\gamma}{\sqrt{n}} \right) + \log\left(1 + \sqrt{e}\right) \right) .
\]

Interestingly, the choice \( Q = Q^0 \) gives the smallest right-hand side in Eq. (6) (however, it does not necessarily minimize the bound on \( Q_S[L] \)) which leads to the following for the Gibbs learner: \( Q_S[L] - Q_S[\hat{L}_S] \leq 1/\sqrt{n} + \gamma/n \). Notice that this bound has an additive \( 1/\sqrt{n} \) compared to the bound in expectation of Raginsky et al. [2017]. See also Catoni [2007, Corollary 1.3.2].

\(^8\)Germain et al. [2009] presented a similar generic PAC-Bayes theorem but with fixed ‘data-free’ priors.

\(^9\)Audibert and Bousquet [2007] separately analyzed the exponential moment but under ‘data-free’ priors.
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 $Q^*$ ∈ $\mathcal{M}_1(\mathcal{H})$ and stochastic kernel $Q^0$ ∈ $\mathcal{K}(\mathcal{S}, \mathcal{H})$ satisfying the DP($\epsilon$) property\(^\text{10}\),

$$\xi(Q^0) \leq 2 \max\{\xi(Q^*), 1\} \exp\left(\frac{n\epsilon^2}{2} + \epsilon \sqrt{\frac{n}{2} \log \left(\frac{2}{\beta}\right)}\right) \quad \beta \in (0, 1). \quad (7)$$

Eq. (7) 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 to different forms of $\xi(Q^*)$—essentially, upper bounds on the exponential moment typically considered in the PAC-Bayesian literature. For example, taking $F(x, y) = n \log(x||y)$ one can show that $\xi(Q^*) \leq 2\sqrt{n}$ [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:

$$\text{kl}(Q|Q^0) \leq \frac{1}{n} \left(\text{KL}(Q||Q^0) + \frac{1}{2}n\epsilon^2 + \epsilon \sqrt{\frac{\log(3\sqrt{n})}{2} n + \log(\frac{3\sqrt{n}}{\delta})}\right).$$

A more general version of Eq. (7), whose derivation is based on the notion of max-information [Dwork et al., 2015a], is discussed in Appendix C (see Lemma 7 there) and proved in Appendix C.1. The details of the conversion recipe are also in Appendix C.

2.3 A generalization bound for the square loss with a data-dependent prior

Our third and last contribution is a novel bound for the setting of learning linear predictors with the square loss. This will demonstrate the full power of our take on the PAC-Bayes analysis, as we will consider a regression problem with the squared loss and a data-dependent prior distribution. In fact, our framework of data-dependent priors makes it possible to obtain the problem-dependent bound in Eq. (8) for square loss regression. We are not aware of an equivalent previous result.

In this setting, the input space is $\mathcal{X} = \mathbb{R}^d$ and the label space $\mathcal{Y} = \mathbb{R}$. A linear predictor is of the form $h_w : \mathbb{R}^d \rightarrow \mathbb{R}$ with $h_w(x) = w^\top x$ for $x \in \mathbb{R}^d$, where of course $w \in \mathbb{R}^d$. Hence $h_w$ may be identified with the weight vector $w$ and correspondingly the hypothesis space $\mathcal{H}$ may be identified with the weight space $W = \mathbb{R}^d$. The size-$n$ random sample is $S = ((X_1, Y_1), \ldots, (X_n, Y_n)) \in (\mathbb{R}^d \times \mathbb{R})^n$. The population and empirical losses are defined with respect to the square loss function:

$$L(w) = \frac{1}{2} \mathbb{E}[(w^\top X - Y)^2] \quad \text{and} \quad \hat{L}_S(w) = \frac{1}{2n} \sum_{i=1}^{n} (w^\top X - Y_i)^2$$

The population covariance matrix is $\Sigma = \mathbb{E}[X_1X_1^\top] \in \mathbb{R}^{d \times d}$ and its eigenvalues are $\lambda_1 \geq \cdots \geq \lambda_d$. The (regularized) sample covariance matrix is $\Sigma_\lambda = (X_1X_1^\top + \cdots + X_nX_n^\top)/n + \lambda I$ for $\lambda > 0$, with eigenvalues $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_d$.

Consider the prior $Q^0_{0, \lambda}$ with density $q^0_{0, \lambda}(w) \propto e^{-\frac{\gamma}{2\lambda}||w||^2}$ for some $\gamma, \lambda > 0$, that possibly depend on the data. In this setting, we prove (Appendix D) that for any posterior $Q \in \mathcal{K}(\mathcal{S}, \mathcal{W})$, for any $\gamma > 0$, and any $\lambda > \max_i \{\lambda_i - \lambda_i\}$, with probability one over size-$n$ random samples $S$ we have

$$Q^0_{S, \lambda} - Q^0_{S, \hat{L}_S} \leq \min_{w \in \mathbb{R}^d} L(w) + \frac{1}{\gamma} \text{KL}(Q || Q^0_{\gamma, \lambda}) + \frac{1}{2\gamma} \sum_{i=1}^{d} \log \left(\frac{\lambda}{\lambda + \tilde{\lambda}_i - \lambda_i}\right). \quad (8)$$

A straightforward observation is that this generalization bound holds with probability one over the distribution of size-$n$ random samples. This is a stronger result than usual high-probability bounds. Of course one may derive a high-probability bound from Eq. (8) by an application of Markov’s inequality, but that would make the result weaker. The stronger result with probability one, for instance, allows to select the best out a countable collection of $\lambda$ values at no extra cost, while the high-probability bound would need to pay a union bound price for such selection.

\(^{10}\)See Appendix C for details on the DP($\epsilon$) property.
Notice that we are not necessarily assuming bounded inputs or labels. Our bound depends on the data-generating distribution (possibly of unbounded support) via the spectra of the covariance matrices. While this is apparent by looking at the last term in Eq. (8), in fact the $\text{KL}(\text{Posterior}||\text{Prior})$ term also depends on the covariances (see Proposition 12 in Appendix D). In particular, if the data inputs are independent sub-gaussian random vectors, then with high probability $|\lambda_i - \lambda_j| \lesssim \sqrt{d/n}$ and the last term in Eq. (8) then behaves as $d \log \left( \frac{\lambda}{(\lambda + \lambda_i - \lambda_j)} \right) \lesssim d/\sqrt{n - 1}$. This of course can be extended to heavy-tailed distributions or, in general, to any input distributions such that spectrum of the covariance matrix concentrates well [Vershynin, 2011].

An important component of the proof of Eq. (8) is the following identity for the exponential moment of $f = \gamma (L(w) - \hat{L}_S(w))$ under the prior distribution: for $\lambda > \max_i \{\lambda_i - \lambda_j\}$, with probability one over random samples $S$,

$$\log Q^0_{\gamma \lambda} [e^f] = \gamma \min_{w \in \mathbb{R}^d} \left( L(w) - (\hat{L}_S(w) + \frac{\lambda}{2} \|w\|_2^2) \right) + \frac{1}{2} \sum_{i=1}^d \ln \left( \frac{\lambda}{\lambda_i - \lambda} \right).$$

This identity computes explicitly the exponential moment of $f$ under the prior distribution. Also this explains why the upper bound in Eq. (8) contains the term $\min_{w \in \mathbb{R}^d} L(w)$. The latter should be understood as the label noise. This term will disappear in a noise-free problem, while given a distribution-dependent boundedness of the loss function, the term will concentrate well around zero (see Proposition 11 in Appendix D). We comment on the free parameter $\gamma$ in Appendix D.

3 Our PAC-Bayes theorem for stochastic kernels

The following results involve data- and hypothesis-dependent functions $f : S \times \mathcal{H} \to \mathbb{R}$. Notice that the order $S \times \mathcal{H}$ is immaterial—functions $\mathcal{H} \times S \to \mathbb{R}$ are treated the same way. It will be convenient to define $f_s(h) = f(s, h)$. If $\rho \in \mathcal{M}_1(\mathcal{H})$ is a ‘data-free’ distribution, we will write $\rho[f_s]$ to denote the $\rho$-average of $f_s(\cdot)$ for fixed $s$, that is, $\rho[f_s] = \int_{\mathcal{H}} f_s(h) \rho(\text{d}h)$. When $\rho$ is data-dependent, that is, $\rho \in \mathcal{K}(S, \mathcal{H})$ is a stochastic kernel, we will write $\rho_s$ 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 $\rho_s[f_s] = \int_{\mathcal{H}} f_s(h) \rho_s(\text{d}h)$.

The joint distribution over $S \times \mathcal{H}$ defined by $P \in \mathcal{M}_1(S)$ and $Q \in \mathcal{K}(S, \mathcal{H})$ is the measure denoted$^{11}$ by $P \otimes Q$ that acts on functions $\phi : S \times \mathcal{H} \to \mathbb{R}$ as follows:

$$(P \otimes Q)[\phi] = \int_S P(\text{d}s) \int_{\mathcal{H}} Q(s, \text{d}h)[\phi(s, h)].$$

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 $E$ denoting the expectation under the joint distribution $P \otimes Q$, the previous display takes the form $E[\phi(S, H)] = E[E[\phi(S, H) | S]]$. Our basic result is the following theorem.

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

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

(i) For any $Q \in \mathcal{K}(S, \mathcal{H})$, 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

$$f(S, H) \leq \log \left( \frac{dQ_S}{d\mathbb{P}_S}(H) \right) + \log(\xi/\delta).$$

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

$$Q_S[f_S] \leq \text{KL}(Q_S||Q^0_S) + \log(\xi/\delta).$$

$^{11}$The notation $P \otimes Q$ (see e.g. Kallenberg [2017]), used here for the joint distribution over $S \times \mathcal{H}$ defined by $P \in \mathcal{M}_1(S)$ and $Q \in \mathcal{K}(S, \mathcal{H})$, corresponds to what in Bayesian learning is commonly written $Q_{H|S}P_S$. 

6
To the best of our knowledge, this theorem is new. Notice that \( Q^0 \) is by default a stochastic kernel from \( \mathcal{S} \) to \( \mathcal{H} \), i.e., a data-dependent distribution over hypotheses. By contrast, the usual PAC-Bayes approaches assume that \( Q^0 \) is a ‘data-free’ distribution, which corresponds to a constant kernel. Also note that the function \( f \) may have unbounded range. A key step of the proof involves a well-known 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:

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

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 \( H \sim Q^0 \) we then have \( \xi = \mathbb{E}^0[e^{f(S,H)}|S] \).

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 \( \xi = \mathbb{E}[e^{f(S,H)}] = \mathbb{E}[e^{D}] \) with \( D = f(S, H) = f(S, H) - \log \left( \frac{dQ}{dQ^0}(H) \right) \).

(i) Applying inequality (**) 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) Recall \( f_S(H) = f(S, H) \). Notice that \( \mathbb{E}[D|S] = Q_S[f_S] - \text{KL}(Q_S||Q_S^0) \). By Jensen inequality, \( \mathbb{E}[D|S] \leq \log \mathbb{E}[e^{D}|S] \). While from (**) applied to \( Y = \mathbb{E}[e^{D}|S] \), with probability at least \( 1 - \delta \) over the random draw of \( S \sim P \) we have \( \log \mathbb{E}[e^{D}|S] \leq \log \mathbb{E}[e^{D}] + \log(1/\delta) \). \qed

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

**Theorem 2 (PAC-Bayes for stochastic kernels)** 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 \( A : \mathcal{S} \times \mathcal{H} \rightarrow \mathbb{R}^k \) and convex function \( F : \mathbb{R}^k \rightarrow \mathbb{R} \), let \( f = F \circ A \) and let \( \xi = (P \otimes Q^0)[e^{f}] \) as in Theorem 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

\[
F(Q_s[A_s]) \leq \text{KL}(Q_s||Q_S^0) + \log(\xi/\delta).
\]

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 \( \xi \) 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 may rewrite \( \xi \) as \( \xi = \mathbb{E}^0[e^{f(S,H)}|S] \) with randomly drawn \( S \sim P \) and \( H \sim Q_S^0 \).

This theorem is a general template for deriving PAC-Bayes bounds (cf. Germain et al. [2009]) not just with ‘data-free’ priors, but also more generally with data-dependent priors. We emphasize that a ‘data-free’ distribution is equivalent to a constant stochastic kernel: \( Q_s^0 = Q_s^0 \) for all \( s, s' \in \mathcal{S} \). Hence \( \mathcal{M}_1(\mathcal{H}) \subset \mathcal{K}(\mathcal{S}, \mathcal{H}) \), but the reverse inclusion is false. To the best of our knowledge, ours is the first work to extend the PAC-Bayes analysis to stochastic kernels.

In contrast to the existing literature on PAC-Bayes bounds, which required ‘data-free’ priors, our Theorem 2 allows the distribution \( Q^0 \) to be data-dependent by default. This is because Theorem 2 holds more generally when \( Q^0 \) is a stochastic kernel. Since ‘data-free’ distributions are constant kernels, the usual cases are encompassed by Theorem 2. The requirement that \( Q^0 \) does not depend on data, as in the literature, plays a role only in the technique used for controlling the exponential moment \( \xi \). This is because with a data-free \( Q^0 \) we may swap the order of integration:

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

Then bounding \( \xi \) proceeds by calculating or bounding \( \xi_{\text{swap}} \) 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 of bounded range played a role when calculating the exponential moment (ξ) term, hence the restriction of bounded loss function can be removed as long as there is a way to bound the exponential moment. This observation may have implications for PAC-Bayesian analyses of learning algorithms under the square loss or the cross-entropy loss, which are unbounded.

As we discussed in Section 2, our Theorem 2 encompasses the usual PAC-Bayes bounds in the literature, but also our theorem allowed to derive novel bounds with data-dependent priors. 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. Importantly, our work highlights that bounding an exponential moment term is where those two usual restrictions played a role, but the restrictions can be relaxed or removed. While the specific PAC-Bayes bound presented here are special cases, deriving additional cases is the topic of ongoing research.

4 Additional discussion and related literature

The literature on PAC-Bayes learning is vast. The usual references are McAllester [1999], Langford and Seeger [2001], and Catoni [2007]; but see also McAllester [2003], Keshet et al. [2011], McAllester [2013], and van Erven [2014]. General forms of the PAC-Bayes theorem have been given before by Audibert [2004] and Germain et al. [2009], and by Bégin et al. [2014, 2016]. Note that McAllester [1999] took further the work of McAllester [1998] which was inspired by the analysis of a Bayesian-style estimator of Shawe-Taylor and Williamson [1997].

There are many application areas that have used the PAC-Bayes approach, but the ways to utilize PAC-Bayes bounds essentially fall into two categories: either use a PAC-Bayes bound to give a risk certificate for a randomized predictor learned by some method, or turn a PAC-Bayes bound itself into a learning method by searching a randomized predictor that minimizes the bound. The latter is mentioned already in McAllester [1999], credit for this approach in various contexts is due also to Germain et al. [2009], Seldin and Tishby [2010], Keshet et al. [2011], Noy and Crammer [2014], Keshet et al. [2017], among others possibly. Subsequent use of this approach for training neural networks was done by Dziugaite and Roy [2017, 2018b]. In fact, the recent resurgence of interest in the PAC-Bayes approach has been to a large extent motivated by the interest in generalization guarantees for neural networks. Langford and Caruana [2001] used McAllester [1999]’s classical PAC-Bayesian bound to evaluate the error of a (stochastic) neural network classifier. Dziugaite and Roy [2017] obtained numerically non-vacuous generalization bounds by optimizing the same bound. Subsequent studies (e.g. Rivasplata et al. [2019], Mhammedi et al. [2019]) continued this approach, sometimes with links to the generalization of stochastic optimization methods (e.g. London [2017], Neyshabur et al. [2018], Dziugaite and Roy [2018a]) or algorithmic stability.

Our work mainly contributes in the direction of connecting PAC-Bayes priors to data. 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, \( Q_S[L] - \inf_{h \in H} L(h) \), 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.

There is also a line of work related to relaxing the restriction of bounded range in the loss functions. A straightforward way to extend results to unbounded losses is to make assumptions on the tail behaviour of the loss [Alquier et al., 2016, Germain et al., 2016] or its moments [Alquier and Guedj, 2018, Holland, 2019], leading to interesting bounds in special cases. An alternative approach was explored by Kuzborskij and Szepesvári [2019], where instead of boundedness, the control of the higher-order moments of the loss is captured by the Efron-Stein variance proxy. Squared loss regression was studied by Shalaeva et al. [2020] which improved results of Germain et al. [2016] and relaxed the data-generation assumption to non-iid data. Or results contribute to this line of work with a particular bound for the square loss that holds without those assumptions.
References

P. Alquier and B. Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107 (5):887–902, May 2018.

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.

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-taylor. Tighter PAC-Bayes bounds. In *Neural Information Processing Systems (NIPS)*, pages 9–16, 2007.

J-Y Audibert. A Better Variance Control For PAC-Bayesian Classification. Preprint, 2004.

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

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian theory for transductive learning. In *Artificial Intelligence and Statistics*, pages 105–113, 2014.

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444, 2016.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

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 Learning (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 *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 Learning (ICML)*, pages 353–360. ACM, 2009.

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Neural Information Processing Systems (NIPS)*, pages 1884–1892, 2016.
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 Neural Information Processing Systems (NIPS), 2019. To appear.

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

J. Keshet, D. McAllester, and T. Hazan. PAC-Bayesian approach for minimization of phoneme error rate. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2224–2227. IEEE, 2011.

J. Keshet, S. Maji, T. Hazan, and T. Jaakkola. Perturbation Models and PAC-Bayesian Generalization Bounds. In Perturbations, Optimization, and Statistics, pages 289–309. MIT Press, 2017.

I. Kuzborskij and C. Szepesvári. Efron-Stein PAC-Bayesian Inequalities. arXiv:1909.01931, 2019. URL https://arxiv.org/abs/1909.01931.

I. Kuzborskij, N. Cesa-Bianchi, and C. Szepesvári. Distribution-Dependent Analysis of Gibbs-ERM Principle. In A. Beygelzimer and D. Hsu, editors, Conference on 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 Neural Information Processing Systems (NIPS), pages 809–816, 2001.

J. Langford and M. Seeger. Bounds for averaging classifiers. Technical Report CMU-CS-01-102, Carnegie Mellon University, 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 Neural Information Processing Systems (NIPS), pages 2931–2940, 2017.

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

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

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

D. A. McAllester. PAC-Bayesian model averaging. In Conference on Learning Theory (COLT), pages 164–170. ACM, 1999.

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

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. Noy and K. Crammer. Robust forward algorithms via PAC-Bayes and Laplace distributions. In Artificial Intelligence and Statistics, pages 678–686, 2014.
E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13(Dec):3507–3531, 2012.

A. Pentina and C. H. Lampert. A PAC-Bayesian Bound for Lifelong Learning. In *International Conference on Machine Learning (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 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 *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.

Y. Seldin and N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11(Dec):3595–3646, 2010.

V. Shalaeva, A. F. Esfahani, P. Germain, and M. Petreczky. Improved PAC-Bayesian Bounds for Linear Regression. In *Conference on Artificial Intelligence (AAAI)*, 2020.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayesian estimator. In *Conference on Learning Theory (COLT)*, pages 2–9. ACM, 1997.

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

I. O. Tolstikhin and Y. Seldin. PAC-Bayes-empirical-Bernstein inequality. In *Neural Information Processing Systems (NIPS)*, pages 109–117, 2013.

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

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2011.
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 \(\nu : \Sigma_\mathcal{X} \rightarrow [0, +\infty]\) such that \(\nu(\emptyset) = 0\). We write \(M(\mathcal{X}, \Sigma_\mathcal{X})\) for the set of all measures on this space, and \(M_1(\mathcal{X}, \Sigma_\mathcal{X})\) 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 \(M(\mathcal{X})\) and \(M_1(\mathcal{X})\), respectively. For any measure \(\nu \in M(\mathcal{X})\) and measurable function \(f : \mathcal{X} \rightarrow \mathbb{R}\), we write \(\nu[f]\) to denote the \(\nu\)-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 \(P \in M_1(\mathcal{X})\), then the expected value of \(f(X)\) is \(E[f(X)] = P[f]\), and its variance is \(\text{Var}[f(X)] = P[f^2] - P[f]^2\).

B 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 \(S \times \mathcal{H}\) defined by \(P \in M_1(S)\) and \(Q \in K(S, \mathcal{H})\). Drawing a random pair \((S, H) \sim P \otimes Q\) is equivalent to drawing \(S \sim P\) and drawing \(H \sim Q_S\). With \(E\) denoting expectation under \(P \otimes Q\), for measurable functions \(\phi : S \times \mathcal{H} \rightarrow \mathbb{R}\) we have \(E[\phi(S, H)] = E[E[\phi(S, H)|S]]\). Also recall \(S = \mathbb{Z}^n\).

Lemma 3 For any \(n\), for any loss function with range \([0, b]\), for any \(Q \in K(S, \mathcal{H})\) such that \(Q_s(dh) \propto e^{-\gamma L(h, s)} \mu(dh)\), the following upper bound on \(\xi(Q) = E[e^{\gamma(L(H) - \hat{L}(H, S))}]\) holds:

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

For the proof of Lemma 3, we will use the shorthand \(\Delta_{\alpha}(h) = \sqrt{n}(L(h) - \hat{L}(h, s))\) where \((s, h) \in S \times \mathcal{H}\). We need two technical results, quoted next for convenience.

Lemma 4 (Boucheron et al. 2013, Lemma 4.18) Let \(Z\) be a real-valued integrable random variable such that

\[
\log E \left[e^{\alpha(Z - E[Z])}\right] \leq \frac{\alpha^2 \sigma^2}{2} \quad (\forall \alpha > 0)
\]

\(^{12}\) For sets \(A \in \Sigma_\mathcal{X}\) the event that the value of \(X\) falls within \(A\) has probability \(P[X \in A] = P(A)\).
holds for some $\sigma > 0$, and let $Z'$ be another real-valued integrable random variable. Then we have $\mathbb{E}[Z'] - \mathbb{E}[Z] \leq \sqrt{2\sigma^2 \text{KL} (\text{Law}(Z') || \text{Law}(Z))}$.

**Lemma 5 (Kuzborskij et al. 2019, Lemma 9)** Let $f_A, f_B : \mathcal{H} \to \mathbb{R}$ be measurable functions such that the normalizing factors
\[
N_A = \int_{\mathcal{H}} e^{-\gamma f_A(h)} \, dh \quad \text{and} \quad N_B = \int_{\mathcal{H}} e^{-\gamma f_B(h)} \, dh
\]
are finite for all $\gamma > 0$, and let $p_A$ and $p_B$ be the corresponding densities:
\[
p_A(h) = \frac{1}{N_A} e^{-\gamma f_A(h)}, \quad p_B(h) = \frac{1}{N_B} e^{-\gamma f_B(h)}, \quad h \in \mathcal{H}.
\]
Whenever $N_A > 0$ we have that
\[
\ln \left( \frac{N_B}{N_A} \right) \leq \gamma \int_{\mathcal{H}} p_B(h) (f_A(h) - f_B(h)) \, dh.
\]

The last lemma is helpful for bounding the log-ratio of Gibbs integrals. The notation ‘$dh$’ 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 3] 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
\[
\mathbb{E} \left[ e^{\Delta_S(H)} \right] \left( \ln \mathbb{E} \left[ e^{\Delta_S(H)} \right] - \ln \mathbb{E} \left[ e^{(\Delta_S(H'))_+} \right] \right) \leq \mathbb{E} \left[ e^{\Delta_S(H)} (\Delta_S(H) - (\Delta_S(H'))_+) \right]
\]
so then rearranging
\[
\ln \mathbb{E} \left[ e^{\Delta_S(H)} \right] \leq \mathbb{E} \left[ \frac{e^{\Delta_S(H)}}{\mathbb{E} \left[ e^{\Delta_S(H)} \right]} (\Delta_S(H) - (\Delta_S(H'))_+) \right] + \ln \mathbb{E} \left[ e^{(\Delta_S(H'))_+} \right]
\]
\[
\leq \mathbb{E} \left[ \frac{e^{\Delta_S(H)}}{\mathbb{E} \left[ e^{\Delta_S(H)} \right]} \Delta_S(H) \right] + \ln \mathbb{E} \left[ e^{(\Delta_S(H'))_+} \right]. \tag{10}
\]

Let’s write $q_s$ for the density of $Q_s$ with respect to a reference measure $dh$ over $\mathcal{H}$, and introduce a measure
\[
d\mu_S(h) = \frac{e^{\Delta_S(h)}}{\mathbb{E} \left[ e^{\Delta_S(h)} \right]} \, dq_S(h), \quad h \in \mathcal{H}.
\]
Then the inequality (10) can be written as
\[
\ln \mathbb{E} \left[ e^{\Delta_S(H)} \right] \leq \mathbb{E} \left( \int_{\mathcal{H}} \Delta_S(h) \, d\mu_S(h) + \ln \mathbb{E} \left[ e^{(\Delta_S(H'))_+} \right] \right).
\]
Bounding (I).} We handle the first term through the stability analysis of the density $\mu_S$. We will denote by $S(1) = (Z_{1:i-1}, Z_1', Z_{i+1:n})$ the sample obtained from $S = (Z_{1:i-1}, Z_i, Z_{i+1:n})$ when replacing the $i$th entry with an independent copy $Z_1'$. In particular,

$$
\frac{1}{\sqrt{n}} \mathbb{E} \left[ \Delta_S(h) \, d\mu_S(h) \right] = \mathbb{E} \left[ \ell(h, Z_1') \, d\mu_S(h) \right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \ell(h, Z_i) \, d\mu_S(h) \right]
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left( \ell(h, Z_1') - \ell(h, Z_i) \right) \, d\mu_S(h) \right]
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \ell(h, Z_i) \, d\mu_{S(i)}(h) - \ell(h, Z_i) \, d\mu_S(h) \right].
$$

The last equality comes from switching $Z_1'$ and $Z_i$ since these variables are distributed identically. Now we use Lemma 4 with $\mu_{S(i)}$ and $\mu_S$, and with $\sigma = b$, to get that

$$
\int \ell(h, Z_i) \, d\mu_{S(i)}(h) - \int \ell(h, Z_i) \, d\mu_S(h) \leq \sqrt{2b^2 \text{KL}(\mu_{S(i)} \parallel \mu_S)}.
$$

Notice that we may use $\sigma = b$ in Lemma 4 since the loss function has range $[0, b]$. Focusing on the KL-divergence, and writing ‘$d\ell$’ for a reference measure on $\mathcal{H}$ with respect to which $q_S, \mu_S, \mu_{S(i)}$ are absolutely continuous,

$$
\text{KL}(\mu_{S(i)} \parallel \mu_S) = \int \ln(d\mu_{S(i)}(h)/d\ell) \, d\mu_{S(i)}(h) - \int \ln(d\mu_S(h)/d\ell) \, d\mu_S(h)
$$

$$
= \int \ln \left( \frac{e^{\Delta_{S(i)}(h)} \cdot e^{-\gamma \tilde{L}_{S(i)}(h)}}{\mathbb{E}[e^{\Delta_S(h)}] \cdot N_{S(i)}} \right) \, d\mu_{S(i)}(h) - \int \ln \left( \frac{e^{\Delta_S(h)} \cdot e^{-\gamma \tilde{L}_{S(i)}(h)}}{\mathbb{E}[e^{\Delta_S(h)}] \cdot N_{S}} \right) \, d\mu_S(h)
$$

$$
= \int \left( \Delta_{S(i)}(h) - \Delta_S(h) \right) \, d\mu_{S(i)}(h) + \ln \left( \frac{N_S}{N_{S(i)}} \right) + \gamma \int \left( \tilde{L}_{S}(h) - \tilde{L}_{S(i)}(h) \right) \, d\mu_{S(i)}(h)
$$

$$
\leq \sqrt{n} \int \left( \tilde{L}_{S}(h) - \tilde{L}_{S(i)}(h) \right) \, d\mu_{S(i)}(h) + \gamma \int \left( \tilde{L}_{S}(h) - \tilde{L}_{S(i)}(h) \right) \, d\mu_S(h)
$$

$$
+ \gamma \int \left( \tilde{L}_{S}(h) - \tilde{L}_{S(i)}(h) \right) \, d\mu_S(h)
$$

$$
= \frac{1}{\sqrt{n}} \int \left[ \ell(h, Z_i) - \ell(h, Z_i') \right] \, d\mu_S(h)
$$

$$
+ \frac{\gamma}{n} \int \left[ \ell(h, Z_i) - \ell(h, Z_i') \right] \, d\mu_S(h)
$$

$$
+ \frac{\gamma}{n} \int \left( \ell(h, Z_i) - \ell(h, Z_i') \right) \, d\mu_{S(i)}(h),
$$

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

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

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

$$
\mathbb{E} \left[ \int \left( \ell(h, Z_i') - \ell(h, Z_i) \right) \, d\mu_S(h) \right] = \mathbb{E} \left[ \int \ell(h, Z_i) \, d\mu_{S(i)}(h) - \int \ell(h, Z_i) \, d\mu_S(h) \right]
$$

$$
\leq \mathbb{E} \left[ \sqrt{2b^2 \text{KL}(\mu_{S(i)} \parallel \mu_S)} \right] \leq \sqrt{2b^2 \mathbb{E}[\text{KL}(\mu_{S(i)} \parallel \mu_S)]}
$$

(By Lemma 4 and Jensen)

$$
= \sqrt{2b^2 \left( \frac{1}{\sqrt{n}} + \frac{2\gamma}{n} \right) \mathbb{E} \left[ \int \left( \ell(h, Z_i') - \ell(h, Z_i) \right) \, d\mu_S(h) \right]}.
$$
The last calculation implies

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

Finally, combining this with Eq. (11) gives

\[ E \int \Delta_S(h) \, d\mu_S(h) \leq 2b^2 \left( 1 + \frac{2\gamma}{\sqrt{n}} \right). \tag{12} \]

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

\[
\ln E \left[ e^{(\Delta_S(H'))_+} \right] = \ln E E \left[ e^{(\Delta_S(H'))_+} | S \right]
= \ln E E \left[ e^{(\Delta_S(H'))_+} | H' \right] \quad \text{(swapping the order of integration)}
\]

and observe that the internal expectation is bounded as

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

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 3.

Using Lemma 3 to bound \( \log(\xi(Q^0)) \) we obtain the following generalization bound by observing that the Gibbs distribution \( Q^0 \) with density \( e^{-\gamma L(h,s)} \) satisfies the DP\((2\gamma/n)\) property.

Corollary 6 For any \( n \), for any \( P_1 \in \mathcal{M}_1(Z) \), for any loss function with range \([0, 1]\), for any \( \gamma > 0 \), for any \( Q^0 \in \mathcal{K}(S, H) \) such that \( Q^0 \propto e^{-\gamma L(h,s)} \) for any \( Q \in \mathcal{K}(S, H) \) and \( \delta \) \((0, 1]\), with probability \( \geq 1 - \delta \) over size-\( n \) i.i.d. samples \( S \sim P_1^n \), we have

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

Proof Theorem 6 of McSherry and Talwar [2007] gives that the Gibbs distribution \( Q^0 \propto e^{-\gamma L(h,s)} \) with potential satisfying \( \sup_{s,s'} \sup_{h \in H} \| \hat{L}_S(h) - \hat{L}_{S'}(h) \| \leq 1/n \) for \( s, s' \in \mathcal{S} \) that differ at most in one entry, satisfies DP\((2\gamma/n)\). Combined with Theorem 8, this gives

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

and applying Pinsker’s inequality \( 2(p - q)^2 \leq \text{kl}(p||q) \) we get

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

\[ \leq \sqrt{\frac{KL(Q_S||Q^0_S)}{2n}} + \gamma \sqrt{\frac{\gamma}{n}} + \sqrt{\frac{\log \left( \frac{3\sqrt{n}}{\delta} \right)}{2n}}. \]
The last inequality is due to the sub-additivity of $t \mapsto \sqrt{t}$. 

While the argument based on d-stability (i.e. Corollary 6) gives a result where the order in $\gamma/n$ matches the one in our bound for the empirical Gibbs prior, our analysis offers an alternative proof technique that might be of independent interest.

### C d-stable data-dependent priors and the max-information lemma

Let $\pi \in K(S, \mathcal{H})$ be a stochastic kernel. Recall that $S = \mathbb{Z}^n$ is the space of size-$n$ samples. When we say that $\pi$ satisfies the DP property with $\epsilon > 0$ (written 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 $e^\epsilon$. Thus, the property entails stability of the data-dependent distribution $\pi_s$ 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 $\xi(n) = (P_n \otimes Q^0)(e^\epsilon)$ for given $P_n \in \mathcal{M}_1(S)$ and $Q^0 \in K(S, \mathcal{H})$. In the following we rely on a notion of $\beta$-approximate max-information [Dwork et al., 2015a,b], which in our context is defined as

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

for $S, S'$ independent copies of each other (same distribution). The next lemma, whose proof is in Appendix C.1, generalizes an idea we learned from Dziugaite and Roy [2018b]:

**Lemma 7** (max-information lemma) Fix $n \in \mathbb{N}$, $P_n \in \mathcal{M}_1(S)$, and a function $f : S \times \mathcal{H} \to \mathbb{R}$. Let $\xi(Q^0) = \int \int e^{f(s, h)} Q^0(dh) P_n(ds)$ for $Q^0 \in \mathcal{M}_1(\mathcal{H})$. Then for any $Q^0 \in K(S, \mathcal{H})$ and for any $\beta \in (0, 1)$ the following bound on $\xi(Q^0) = \int \int e^{f(s, h)} Q^0(dh) P_n(ds)$ holds:

$$\xi(Q^0) \leq 2 \max\{\xi(Q^*), 1\} \exp\{I_{\infty}^\beta(S; Q^0_S)\}.$$

In particular, if $\xi(Q^*) \leq \xi_{bd}$, then $\xi(Q^0) \leq 2 \max\{\xi_{bd}, 1\} \exp\{I_{\infty}^\beta(S; Q^0_S)\}$. This 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 $Q \in K(S, \mathcal{H})$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ over size-$n$ random samples $S \sim P_n$ we have

$$F(Q_S[A_S]) \leq KL(Q_S || Q^0) + \log(\xi_{bd}/\delta). \quad (13)$$

This is written in the generic framework of Theorem 2 where $f(s, h) = F(A(s, h))$, and $\xi_{bd}$ is an upper bound on $\xi = \mathbb{E}^0[e^{f(S, H)}]$ valid when $Q^0$ is a data-free distribution. Then by Lemma 7, for any $Q^0, Q \in K(S, \mathcal{H})$, for any $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ over size-$n$ random samples $S \sim P_n$, we have

$$F(Q_S[A_S]) \leq KL(Q_S || Q^0_S) + \log(2 \max\{\xi_{bd}, 1\}/\delta) + I_{\infty}^\beta(S; Q^0_S). \quad (14)$$

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

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

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 $\xi_{bd}$ is available. For instance, starting from the PAC-Bayes-kl bound (Seeger [2002], see also Langford [2005]) we derive the following:
\textbf{Theorem 8} For any \(n\), for any \(P_1 \in \mathcal{M}_1(Z)\), for any \(Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{H})\) satisfying \(\text{DP}(\epsilon)\), 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 \(\geq 1 - \delta\) over size-\(n\) i.i.d. samples \(S \sim P_1^n\) we have

\[
\text{kl}(Q_S^*[L_s]) || Q_S^*[L_i]) \leq \frac{\text{KL}(Q_S^0 || Q_S^0) + \log\left(\frac{24n^2}{n}\right) + \epsilon n^2 + \epsilon \sqrt{\frac{n}{2}} \log(3)}{\epsilon}.
\] (15)

This is essentially equivalent to [Dziugaite and Roy, 2018b, Theorem 4.2] but with slightly improved constants. The proof of Theorem 8 is as follows.

Under the restrictions of the theorem, we may use \(\xi_{bd} = 2\sqrt{n}\) (as per Maurer [2004]) when the prior is a fixed ‘data-free’ distribution. Then by Lemma 7 we get \(\xi(Q^0) \leq 2\sqrt{n}\text{I}_{\infty}^{f}(S; Q_S^0) + \beta\) when \(Q^0\) is data-dependent. Thus \(\xi(Q^0) \leq 3\sqrt{n}\text{I}_{\infty}^{f}(S; Q_S^0)\), which gives

\[
\log(\xi(Q^0)) \leq \log(3\sqrt{n}) + \text{I}_{\infty}^{f}(S; Q_S^0).
\]

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

\[
\text{I}_{\infty}^{f}(S; Q_S^0) \leq \frac{n\epsilon^2}{2} + \epsilon \sqrt{\frac{n}{2}} \log\left(\frac{2}{\beta}\right).
\]

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

\section{C.1 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 \(\xi_{bd}\) is an upper bound on \(\xi(Q^*) = \int \int e^{f(s, h)} Q^*(dh) P_n(ds)\) which is valid when \(Q^* \in \mathcal{M}_1(\mathcal{H})\) is a fixed ‘data-free’ distribution. 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 \(Q_S^0\) over \(\mathcal{H}\). The corresponding \(\beta\)-approximate max-information as defined by Dwork et al. [2015b] (see also Dwork et al. [2015a]) is denoted \(\text{I}_{\infty}^{f}(S; Q_S^0)\) in our context. The max-information argument to bound \(\xi(Q^0)\) goes as follows:

\[
\xi(Q^0) = \int_{\mathcal{S}} \int_{\mathcal{H}} e^{f(s, h)} Q_s^0(dh) P_n(ds)
\leq e^{\text{I}_{\infty}^{f}(S; Q_S^0)} \int_{\mathcal{S}} \int_{\mathcal{H}} e^{f(s, h)} Q_s^0(dh) P_n(ds) P_n(ds') + \beta,
\]

\[
\leq e^{\text{I}_{\infty}^{f}(S; Q_S^0)} \xi_{bd} + \beta.
\]

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

Thus we get \(\xi(Q^0) \leq 2 \max\{\xi_{bd}, 1\} e^{\text{I}_{\infty}^{f}(S; Q_S^0)}\) by considering the cases \(\xi_{bd} \leq 1\) and \(\xi_{bd} > 1\). This finishes the proof of the “max-information lemma” (Lemma 7).

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

\[
\xi(Q^0) = \int_{\mathcal{S}} \int_{\mathcal{H}} e^{f(s, h)} Q_s^0(dh) P_n(ds)
\]

we may change the measure \(Q_s^0\) to \(Q_s^0\), with any fixed \(s' \in \mathcal{S}\), and the Radon-Nikodym derivative satisfies \(dQ_s^0/dQ_s^0 \leq e^{n\epsilon},\) so we have

\[
\xi(Q^0) \leq e^{n\epsilon} \int_{\mathcal{S}} \int_{\mathcal{H}} e^{f(s, h)} Q_s^0(dh) P_n(ds) \leq e^{n\epsilon} \xi_{bd}
\]

where the integral on the right hand side is upper bounded by \(\xi_{bd}\) since \(Q_s^0\) is now a fixed distribution (with respect to the variable \(s\) of the outer integral). Thus the max-information lemma gives a refined analysis leading to an upper bound on \(\xi(Q^0)\) where ‘\(n\epsilon\)’ is replaced with \(\text{I}_{\infty}^{f}(S; Q_S^0)\).
D Proof of the bound for least squares regression

Let us recall the setting. The input space is $\mathcal{X} = \mathbb{R}^d$ and the label space $\mathcal{Y} = \mathbb{R}$. A linear predictor is of the form $h_w : \mathbb{R}^d \to \mathcal{Y}$ with $h_w(x) = w^\top x$ for $x \in \mathbb{R}^d$, where of course $w \in \mathbb{R}^d$. Hence we may identify $h_w$ with $w$ and correspondingly the hypothesis space $\mathcal{H}$ may be identified with the weight space $W = \mathbb{R}^d$. The size-$n$ random sample is $S = ((X_1, Y_1), \ldots, (X_n, Y_n)) \in (\mathbb{R}^d \times \mathbb{R})^n$.

We are interested in the generalization gap $\Delta^S_w = L(w) - \hat{L}_S(w)$, defined for $w \in \mathbb{R}^d$, where

$$L(w) = \frac{1}{2} \mathbb{E}[(w^\top X_1 - Y_1)^2] \quad \text{and} \quad \hat{L}_S(w) = \frac{1}{2n} \sum_{i=1}^n (w^\top X_i - Y_i)^2$$

are, respectively, the population and empirical losses under the square loss function. For $\lambda > 0$, let $L_{S, \lambda}(w) = \hat{L}_S(w) + (\lambda/2)\|w\|^2$ be the regularized empirical loss, and $\Delta_{S, \lambda}^S = L(w) - L_{S, \lambda}(w)$.

The population covariance matrix is $\Sigma = \mathbb{E}[X_1 X_1^\top] \in \mathbb{R}^{d \times d}$ and its eigenvalues are $\lambda_1 \geq \cdots \geq \lambda_d$. The (regularized) sample covariance matrix is $\hat{\Sigma}_d = (X_1 X_1^\top + \cdots + X_n X_n^\top)/n + \lambda I$ for $\lambda > 0$, with eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$.

By the well-known change-of-measure (Csiszár [1975], Donsker and Varadhan [1975]), for any ('prior') density $q^0$ the following holds:

$$\int_{\mathbb{R}^d} \Delta_{S, \gamma}^S q_S(w) \, dw \leq \text{KL}(q_S \| q^0) + \ln \int_{\mathbb{R}^d} e^{\Delta_{S, \gamma}^S} q^0(w) \, dw. \quad (16)$$

Note that for simplicity we are saying ‘density $p(w)$’ when in fact what we have in mind is that $p$ is the Radon-Nikodym derivative of a probability $P \in \mathcal{M}_1(\mathbb{R}^d)$ with respect to Lebesgue measure, i.e. $P(A) = \int_A p(w) \, dw$ for Borel sets $A \subset \mathbb{R}^d$.

The main theorem and its proof are as follows. Note that this theorem provides a bound on expected generalization gap, which holds with probability one.

**Theorem 9** For any probability kernel $q$ from $S$ to $\mathbb{R}^d$, for any $\gamma > 0$ and $\lambda > \max_i \{\lambda_i - \hat{\lambda}_i\}$, with probability one over random samples $S$,

$$\int_{\mathbb{R}^d} \Delta_{S, \gamma}^S q_S(w) \, dw \leq \min_{w \in \mathbb{R}^d} \Delta_{S, \gamma}^{S, \lambda} + \frac{1}{\gamma} \text{KL}(q_S \| q^0) + \frac{1}{2\gamma} \sum_{i=1}^d \ln \left( \frac{\lambda}{\lambda + \hat{\lambda}_i - \lambda_i} \right).$$

**Proof** We get the statement by combining Eq. (16) with the analytic form of exponential moment of $\gamma \Delta_{S, \gamma}^S$ given by Lemma 10 below.

**Lemma 10 (exponential moment)** Let $q^0(w) \propto e^{-\gamma \lambda/\|w\|^2}$ for $\gamma > 0$ and $\lambda > \max_i \{\lambda_i - \hat{\lambda}_i\}$. Then, with probability one over random samples $S$,

$$\ln \int_{\mathbb{R}^d} e^{\gamma \Delta_{S, \gamma}^S} q^0(w) \, dw = \gamma \min_{w \in \mathbb{R}^d} \Delta_{S, \gamma}^{S, \lambda} + \frac{1}{2} \sum_{i=1}^d \ln \left( \frac{\lambda}{\lambda + \hat{\lambda}_i - \lambda_i} \right).$$

This lemma fills in the main part of the proof of Theorem 9. Notice that this lemma computes explicitly the exponential moment of $\gamma \Delta_{S, \gamma}^S$, without making additional assumptions on the loss function. The proofs of this lemma and of other results in this section are deferred to Appendix D.1.

A couple of comments about Theorem 9. First, note that the inequality holds almost surely (a.s.) over samples $S$ which differs from the usual PAC-Bayesian analysis because we did not apply Markov inequality. However, one can still convert the bound we obtained above to a high-probability bound, by looking at the concentration of eigenvalues of the sample covariance matrix (which will require appropriate assumptions on the marginal distribution). Second, we have a new term $\min_{w \in \mathbb{R}^d} \sum_{i=1}^d \Delta_{S, \gamma}^{S, \lambda}$ whose range is directly connected to that of the loss function. This term is problem-dependent. Indeed, the following straightforward proposition lets us understand better its role.
Proposition 11 (regularized gap) If \( w^* \in \arg \min_{w \in \mathbb{R}^d} L(w) \), so that \( L(w^*) = \min_{w \in \mathbb{R}^d} L(w) \), then with probability one over random samples \( S \) we have that

\[
\min_{w \in \mathbb{R}^d} \Delta_{w}^{S,\lambda} \leq L(w^*) .
\]

If \( \max_i (X_i^T w^* - Y_i)^2 \leq B \) a.s., then for any \( x > 0 \), with probability at least \( 1 - e^{-x} \) we have that

\[
\min_{w \in \mathbb{R}^d} \Delta_{w}^{S,\lambda} \leq B \sqrt{\frac{x}{2n}} .
\]

The first part of Proposition 11 implies that in a noise-free problem the term \( \min_{w \in \mathbb{R}^d} \Delta_{w}^{S,\lambda} \) will disappear; while the second part argues that given a distribution-dependent boundedness of the loss function, the term will concentrate well around zero.

Now we turn our attention to the KL(Posterior||Prior) term, stated analytically by the following proposition:

Proposition 12 (KL term) For \( q_S(w) \propto e^{-\frac{1}{2} L_{S,a}(w)} \) and \( q^0(w) \propto e^{-\frac{\alpha}{2} \|w\|^2} \) and any \( \alpha, \lambda, \gamma > 0 \),

\[
\text{KL}(q_S \| q^0) = \frac{1}{2} \left( \ln \det \left( \frac{1}{\lambda} \Sigma_n \right) \right) + \text{tr} \left( \lambda \Sigma_n^{-1} - I \right) + \frac{\lambda \gamma}{n^2} \sum_{i=1}^n Y_i^2 \|X_i\|^2 \Sigma_n^{-2} .
\]

Furthermore, if \( \max_i \|X_i\|_2 \leq 1 \) a.s., then

\[
\text{KL}(q_S \| q^0) \leq \frac{1}{2} \left( d \ln \left( \frac{1 + \alpha}{\lambda} \right) + d \left( \frac{\lambda}{\lambda_d + \alpha} - 1 \right) + \frac{\lambda \gamma}{n^2} \sum_{i=1}^n Y_i^2 \|X_i\|^2 \Sigma_n^{-2} \right) .
\]

Combining the results outlined above yields the following corollary.

Corollary 13 (data-dependent bound) Let \( \epsilon_n = \max \{ \lambda_i - \hat{\lambda}_i \} \), and choose \( \lambda = c \epsilon_n \) for some \( c > 1 \). Then, with probability one over random samples \( S \),

\[
\int_{\mathbb{R}^d} \Delta_w^S q_S(w) \, dw \leq \min_{w \in \mathbb{R}^d} \Delta_w^{S,c \epsilon_n} + d \frac{\ln \left( \frac{1 + \alpha}{e(c - 1) \epsilon_n} \right)}{2 \gamma} + \frac{c \epsilon_n d}{\lambda_d + \alpha} \left( \frac{1}{2 \gamma} + \frac{1}{n} \sum_{i=1}^n Y_i^2 \right) .
\]

Finally, a quick comment on the free parameter \( \gamma > 0 \) in our bound of Theorem 9. In the standard PAC-Bayes analysis one would see a trade-off in \( \gamma \), with a usual near-optimal setting of \( \gamma = \sqrt{n} \) [Shalaeva et al., 2020]. Such trade-off is more subtle in our Theorem 9 since one would need to ensure that \( \gamma^{-1} \text{KL}(q_S \| q^0_{\gamma,\lambda}) \to 0 \) as \( \gamma \to \infty \) for the desired choice of \( q_S \).

D.1 Proofs

Proof [Proof of Lemma 10] For convenience we introduce the abbreviations \( s = \mathbb{E}[Y_1 X_1] \) and its empirical counterpart \( \hat{S} = (Y_1 X_1 + \cdots + Y_n X_n)/n \). Also let’s define \( C = \mathbb{E}[Y_i^2] - (Y_1^2 + \cdots + Y_n^2)/n \). The density is \( q^0(w) = Z_0^{-1} e^{-\frac{\alpha}{2} \|w\|^2} \), with \( Z_0 \) a normalizing factor. A straightforward
where Eq. \((\ref{eq:gaussian_integration})\) is just rewriting things, while in Eq. \((\ref{eq:gaussian_integration2})\) we assume that \(\lambda > \max_i \{\lambda_i - \hat{\lambda}_i\}\). Eqs. \((\ref{eq:gaussian_integration2})\) and \((\ref{eq:gaussian_integration3})\) come from Gaussian integration, and Eq. \((\ref{eq:gaussian_integration3})\) is a consequence of:

**Proposition 14** Assuming that \(\lambda > \max_i \{\lambda_i - \hat{\lambda}_i\}\),

\[
\min_{w \in \mathbb{R}^d} \left\{ L(w) - \hat{L}_{S,\lambda}(w) \right\} = C + \frac{1}{2} (s - \hat{S})^\top (\hat{\Sigma}_\lambda - \Sigma)^{-1} (s - \hat{S}) .
\]

Finally, taking logarithm of the integral completes the proof of Lemma 10.  

**Proof** [Proof of Proposition 14] Observe that

\[
\nabla_w \left( c - \frac{1}{2} w^\top (\hat{\Sigma}_\lambda - \Sigma) w - (\hat{S} - s)^\top w \right) = -(\hat{\Sigma}_\lambda - \Sigma) w + (s - \hat{S}) .
\]

For \(\lambda > \max_i \{\lambda_i - \hat{\lambda}_i\}\) the matrix \((\hat{\Sigma}_\lambda - \Sigma)\) is positive definite, and plugging the solution of \(\nabla_w = 0\), namely \(\hat{w} = (\hat{\Sigma}_\lambda - \Sigma)^{-1} (s - \hat{S})\), back into the objective we get

\[
C - \frac{1}{2} \hat{w}^\top (\hat{\Sigma}_\lambda - \Sigma) \hat{w} + (s - \hat{S})^\top w = C + \frac{1}{2} (s - \hat{S})^\top (\hat{\Sigma}_\lambda - \Sigma)^{-1} (s - \hat{S})
\]

which completes the proof of Proposition 14.  

**Proof** [Proof of Proposition 11] Clearly \(\min_{w \in \mathbb{R}^d} \Delta_w^{S,\lambda} \leq \Delta_w^{S,B} \leq L(w^*)\), which proves the first part of the proposition. For the second part, under the assumption that \(\max_i (X_i^\top w^* - Y_i)^2 \leq B\) a.s., Hoeffding’s inequality gives:

\[
\Delta_w^{S,\lambda} \leq \frac{1}{2} \mathbb{E}[(X_i^\top w^* - Y_i)^2] - \frac{1}{2n} \sum_{i=1}^n (X_i^\top w^* - Y_i)^2 \leq B \sqrt{\frac{x}{2n}} .
\]

This completes the proof of Proposition 11.  

**Proof** [Proof of Proposition 12] Observe that

\[
q_S(w) = \frac{e^{-\frac{1}{2} w^\top \hat{\Sigma}_u w + \gamma w^\top S - \frac{1}{2} Y^2}}{\int_{\mathbb{R}^d} e^{-\frac{1}{2} u^\top \hat{\Sigma}_u u + \gamma u^\top S - \frac{1}{2} Y^2} du} = \frac{e^{-\frac{1}{2} w^\top \hat{\Sigma}_u w + \gamma w^\top S - \frac{1}{2} S^\top \hat{\Sigma}_u^{-1} S}}{\int_{\mathbb{R}^d} e^{-\frac{1}{2} u^\top \hat{\Sigma}_u u + \gamma u^\top S - \frac{1}{2} S^\top \hat{\Sigma}_u^{-1} S} du} =: G(w) \alpha e^{-\frac{1}{2} (w - \hat{\Sigma}_u^{-1} S)^\top \hat{\Sigma}_u (w - \hat{\Sigma}_u^{-1} S)} ,
\]
where \( \hat{S} = (Y_1X_1 + \cdots + Y_nX_n)/n \) and \( \hat{Y}^2 = (Y_1^2 + \cdots + Y_n^2)/n \). Recall that analytic form of KL-divergence between two Gaussians is:

\[
\text{KL} \left( \text{Gauss}(x_1, A_1) \mid \mid \text{Gauss}(x_0, A_0) \right) = \frac{1}{2} \left( \ln \left( \frac{\det A_0}{\det A_1} \right) + \text{tr} \left( A_0^{-1}A_1 \right) - d + (x_1 - x_0)^\top A_0^{-1}(x_1 - x_0) \right)
\]

This gives

\[
\text{KL} \left( q_S \mid \mid q^0 \right) = \frac{1}{2} \left( \ln \det \left( \frac{1}{\lambda} \hat{\Sigma}_a \right) + \text{tr} \left( \lambda \hat{\Sigma}_a^{-1} - I \right) + \lambda \gamma \hat{S}^\top \hat{\Sigma}_a^{-2} \hat{S} \right)
\]

This shows the first statement.

The ‘furthermore’ statement is shown using a simple fact that for \( d \times d \) positive definite matrix \( A \), we have

\[
\ln \det \left( \frac{1}{\lambda} \hat{\Sigma}_a \right) \leq d \ln \text{tr} \left( \frac{1}{d\lambda} \hat{\Sigma}_a \right) \leq d \ln \left( \frac{1 + \alpha}{\lambda} \right)
\]

where we have assumed that \( \max_i \|X_i\|_2 \leq 1 \) a.s. and the fact

\[
\text{tr} \left( \lambda \hat{\Sigma}_a^{-1} - I \right) \leq d \left( \frac{\lambda}{\lambda_d + \alpha} - 1 \right).
\]

This completes the proof of Proposition 12.

**Proof** [Proof of Corollary 13] Theorem 9 combined with Proposition 12 gives us

\[
\int_{\mathbb{R}^d} \Delta^S_w q_S(w) \, dw \leq \min_{w \in \mathbb{R}^d} \Delta^S_w + \frac{d}{2\gamma} \ln \left( \frac{1 + \alpha}{\lambda} \right) + \frac{d}{2\gamma} \left( \frac{\lambda}{\lambda_d + \alpha} - 1 \right)
\]

\[
+ \frac{\lambda}{2n^2} \sum_{i=1}^n Y_i^2 \|X_i\|_2^2 \|\hat{\Sigma}_a^{-2}\| - 1 + \frac{d}{2\gamma} \sum_{i=1}^d \ln \left( \frac{\lambda}{\lambda + \lambda_i - \lambda_i} \right)
\]

\[
\leq \min_{w \in \mathbb{R}^d} \Delta^S_{w,c\hat{\varepsilon}_n} + \frac{d}{2\gamma} \ln \left( \frac{1 + \alpha}{c\hat{\varepsilon}_n} \right) + \frac{d}{2\gamma} \left( \frac{c\hat{\varepsilon}_n}{\lambda_d + \alpha} - 1 \right)
\]

\[
+ \frac{cd\hat{\varepsilon}_n}{\lambda_d + \alpha} \left( \frac{1}{n} \sum_{i=1}^n Y_i^2 \right) + \frac{d}{2\gamma} \ln \left( \frac{c}{c - 1} \right)
\]

\[
\leq \min_{w \in \mathbb{R}^d} \Delta^S_{w,c\hat{\varepsilon}_n} + \frac{d}{2\gamma} \ln \left( \frac{1 + \alpha}{c(c - 1)\hat{\varepsilon}_n} \right) + \frac{c\hat{\varepsilon}_n d}{\lambda_d + \alpha} \left( \frac{1}{2\gamma} + \frac{1}{n} \sum_{i=1}^n Y_i^2 \right),
\]

where we used the fact that

\[
\sum_{i=1}^d \ln \left( \frac{\lambda}{\lambda + \lambda_i - \lambda_i} \right) = \sum_{i=1}^d \ln \left( \frac{c \max_i \{\lambda_i - \hat{\lambda}_i\}}{c \max_i \{\lambda_i - \hat{\lambda}_i\} - (\lambda_i - \hat{\lambda}_i)} \right) \leq d \ln \left( \frac{c}{c - 1} \right)
\]

and by a simple SVD argument

\[
\frac{1}{n^2} \sum_{i=1}^n Y_i^2 \|X_i\|_2^2 \leq \frac{d}{n(\lambda_d + \alpha)} \sum_{i=1}^n Y_i^2.
\]

This completes the proof of Corollary 13.