# 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 requirements of fixed 'data-free' priors, bounded losses, and i.i.d. data. We highlight that those requirements were used to upper-bound an exponential moment term, while the basic PAC-Bayes theorem remains valid without those restrictions. We present three bounds that illustrate the use of data-dependent priors, including one for the unbounded square loss.

## 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. $\mathcal{Z} = \mathbb{R}^d \times \mathcal{Y}$ for a supervised learning problem where the input space is $\mathbb{R}^d$ and the label set is $\mathcal{Y}$) 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 *empirical loss* is $Q[\hat{L}_s] = \int_{\mathcal{H}} \hat{L}_s(h) Q(dh)$, i.e. the $Q$-average of the standard empirical loss $\hat{L}_s(h) = \hat{L}(h, s)$ defined as $\hat{L}(h, s) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i)$ for a fixed $h \in \mathcal{H}$ and $s = (z_1, \ldots, z_n)$, where $\ell : \mathcal{H} \times \mathcal{Z} \to [0, \infty)$ is a given loss function. Similarly, the expected *population loss* is $Q[L] = \int_{\mathcal{H}} L(h) Q(dh)$, i.e. the $Q$-average of the standard population loss $L(h) = \int_{\mathcal{Z}} \ell(h, z) P_1(dz)$ for a fixed $h \in \mathcal{H}$, where $P_1 \in \mathcal{M}_1(\mathcal{Z})$ is the distribution that generates one random example.

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 $\mathcal{H}$.

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

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

**Data-dependent distributions as stochastic kernels.** 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[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 $Q_s : B \mapsto Q(s,B)$ is a probability measure over $\mathcal{H}$. We write $\mathcal{K}(\mathcal{S},\mathcal{H})$ to denote the set of all stochastic kernels from $\mathcal{S}$ to —distributions over— $\mathcal{H}$. We reserve the notation $\mathcal{M}_1(\mathcal{H})$ for the set of 'data-free' distributions over $\mathcal{H}$. Notice that $\mathcal{M}_1(\mathcal{H}) \subset \mathcal{K}(\mathcal{S},\mathcal{H})$, since every 'data-free' distribution can be regarded as a constant kernel.

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 $\mu$ a base measure over $\mathcal{H}$. Note that, besides randomized predictors, other prediction schemes may be devised from a learned distribution over hypotheses, as for instance ensemble predictors and majority vote predictors (see the related literature in Section 4 below).

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" [Freund, 1998] or "self-certified" [Pérez-Ortiz et al., 2020] learning algorithms, which are learning algorithms that use all the available data to simultaneously provide a predictor and a risk certificate that is valid on unseen examples.

*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].

In the following, for any given $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ and $s \in \mathcal{S}$, we write $Q_s[\hat{L}_s] = \int \hat{L}_s(h) Q_s(dh)$ and $Q_s[L] = \int L(h) Q_s(dh)$ for the expected empirical loss and the expected population loss, respectively. 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 $S$:

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

$\mathrm{KL}(\cdot\|\cdot)$ stands for the Kullback-Leibler divergence[4] which is defined for two given probability distributions $Q, Q'$ over $\mathcal{H}$ as follows: $\mathrm{KL}(Q\|Q') = \int_{\mathcal{H}} \log\left(dQ/dQ'\right) dQ$, where $dQ/dQ'$ denotes the Radon-Nikodym derivative. Note that PAC-Bayes bounds (e.g. McAllester's bound described above) are usually presented under a statement that says that with probability at least $1-\delta$, the

[2] This is also called a *transition kernel* or *probability kernel*, a well-known concept in the literature on stochastic processes, see e.g. Kallenberg [2017], Meyn and Tweedie [2009], Ethier and Kurtz [1986].

[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 that we denote $\Sigma_{\mathcal{H}}$. For precise definitions see Appendix A.

[4] Also known as relative entropy, see e.g. Cover and Thomas [2006].

displayed inequality holds simultaneously for all probability distributions $Q$ over $\mathcal{H}$, i.e. with an arbitrary $Q$ replacing $Q_S$. Such commonly used formulation has the apparent advantage of being valid uniformly for all distributions over $\mathcal{H}$, while our formulation is valid for a fixed kernel. At the same time, the commonly used formulation has the disadvantage of hiding the data-dependence of the 'posterior' distributions used in practice, while our formulation in terms of a stochastic kernel shows explicitly the data-dependence: given the data $S$, the corresponding distribution over $\mathcal{H}$ is $Q_S$. Notice that one fixed stochastic kernel suffices in order to describe a whole parametric family of distributions (such as Gaussian or Laplace distributions, among others) with parameter values learned from data. Since our main interest is in results for data-dependent distributions (contrasted to results for fixed 'data-free' distributions), we argue in favour of the formulation based on stochastic kernels. These have appeared in the learning theory literature under the names of Markov kernels [Xu and Raginsky, 2017] or regular conditional probabilities [Catoni, 2004, 2007, Alquier, 2008].

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 $Q_S[\hat{L}_S]$. For instance Langford and Seeger [2001] and Seeger [2002] gave an upper bound on the relative entropy of $Q_S[\hat{L}_S]$ and $Q_S[L]$, commonly called the PAC-Bayes-kl bound [Seldin et al., 2012], which holds with high probability over randomly drawn size-$n$ samples $S$:

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

$\mathrm{kl}(\cdot\|\cdot)$, appearing on the left-hand side of this inequality, denotes the binary KL divergence, which is by definition the KL divergence between the Bernoulli distributions with the given parameters:

$$\mathrm{kl}(q\|q') = q\log(\frac{q}{q'}) + (1-q)\log(\frac{1-q}{1-q'}) \quad \text{for } \; q, q' \in [0,1].$$

Inequality (2) is tighter than (1) due to Pinsker's inequality $2(p-q)^2 \leq \mathrm{kl}(p\|q)$. In fact, by a refined form of Pinsker's inequality, namely $(p-q)^2/(2q) \leq \mathrm{kl}(p\|q)$ which is valid for $p < q$ (and tighter than the former when $q < 0.25$), from Eq. (2) one obtains a *localised* inequality[5] (see Eq. (6) of McAllester [2003]), which holds with high probability[6] over randomly drawn size-$n$ samples $S$:

$$Q_S[L] - Q_S[\hat{L}_S] \lesssim \sqrt{\frac{Q_S[\hat{L}_S]}{n}\,\mathrm{KL}(Q_S\|Q^0) + \frac{1}{n}\,\mathrm{KL}(Q_S\|Q^0)} . \tag{3}$$

PAC-Bayes bounds like Eq. (1) and Eq. (3) tell 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. Note that inequality (3) is tighter than (1) when $Q_S[\hat{L}_S] < Q_S[L] < 0.25$. Obviously, the upper bound in Eq. (3) is dominated by the lower-order (second) term whenever the empirical loss $Q_S[\hat{L}_S]$ 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 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 that offers 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 prior distributions to be data-dependent introduces much needed flexibility, since this opens up the possibility to minimize upper bounds in both the posterior *and the prior*, which should lead to tighter empirical bounds on $Q_S[L]$ and tighter risk certificates.

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

[6] The notation $\lesssim$ hides universal constants and logarithmic factors.

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 that were trained on a held-out portion of the available data, thus enabling empirical bounds with PAC-Bayes priors that are 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 $Q_s^0$ (where $Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{H})$) to be stable with respect to 'small' changes in the composition of the $n$-tuple $s$. As we will see shortly, there is benefit in relaxing the restrictions of the usual PAC-Bayes literature.

## 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). The formulation of both these results is based on representing data-dependent distributions as stochastic kernels. To make a case for the usefulness of this approach, we show that our Theorem 2 encompasses many usual bounds which appear in the literature [McAllester, 1998, 1999, Seeger, 2002, Catoni, 2007, Thiemann et al., 2017], while at the same time it enables new PAC-Bayes inequalities. Importantly, our study takes a critical stand on the "usual assumptions" on which PAC-Bayes inequalities are based, namely, (a) data-free prior, (b) bounded loss, and (c) i.i.d. data observations. We aim to clarify the role of these assumptions and to illustrate how to obtain PAC-Bayes inequalities in cases where these assumptions are removed. As we will soon see, the analysis leading to our Theorem 2 shows that the PAC-Bayes priors can be data-dependent by default, and also that the underlying loss function can be unbounded by default. Furthermore, the proof of our Theorem 2 does not rely on the assumption of i.i.d. data observations, which may enable new results for statistically dependent data in future research.

For illustration, our general PAC-Bayes theorem[7] for stochastic kernels (Theorem 2 in Section 3), in specialized form, implies that for any convex function $F : \mathbb{R}^2 \to \mathbb{R}$, for any stochastic kernels $Q, Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over randomly drawn $S$ one has

$$F(Q_S[L], Q_S[\hat{L}_S]) \leq \mathrm{KL}(Q_S \| Q_S^0) + \log(\xi(Q^0)/\delta) \,, \tag{4}$$

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

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

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 [1999]-type bound, $F(x, y) = n\,\mathrm{kl}(y\|x)$ gives the bound of Seeger [2002], and $F(x, y) = n \log\left(\frac{1}{1 - x(1 - e^{-\lambda})}\right) - \lambda n y$ gives the bound of Catoni [2007]. Furthermore, $F(x, y) = n(x - y)^2/(2x)$ leads to the so-called PAC-Bayes-$\lambda$ bound of Thiemann et al. [2017], or to the bound of Rivasplata et al. [2019] which holds under the usual requirements of fixed 'data-free' prior $Q^0$, losses within the $[0, 1]$ range, and i.i.d. data:

$$Q_S[L] \leq \left( \sqrt{Q_S[\hat{L}_S] + \frac{\mathrm{KL}(Q_S \| Q^0) + \log(\frac{2\sqrt{n}}{\delta})}{2n}} + \sqrt{\frac{\mathrm{KL}(Q_S \| Q^0) + \log(\frac{2\sqrt{n}}{\delta})}{2n}} \right)^2 . \tag{5}$$

As consequence of the universality of Eq. (4), besides the usual bounds we may derive novel bounds, e.g. with data-dependent priors $Q_S^0$. 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 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). While previous works[8] analysed separately the exponential moment,

[7] Generic PAC-Bayes theorems, similar in spirit to ours, have been presented before, e.g. by Audibert [2004], Germain et al. [2009], Bégin et al. [2014, 2016], but only with fixed 'data-free' priors.

[8] Audibert and Bousquet [2007], Alquier et al. [2016], among others, for the case of fixed 'data-free' priors.

as we do here, to the best of our knowledge they considered data-free priors only. We think our work is the first to point out techniques to upper bound $\xi(Q^0)$ when $Q^0$ is a stochastic kernel, and to present PAC-Bayesian inequalities where the prior is data-dependent by default. Our work also clarifies where / how the data-free nature of the priors was used in previous works.

We emphasize that in this paper the main focus is on using data-dependent priors 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, nor bounded loss functions nor i.i.d. data observations. The same can be said of Theorem 2, a consequence of Theorem 1(ii), which gives a general template for deriving PAC-Bayes bounds. Below we discuss three generalization bounds with data-dependent priors, two of which are for bounded losses, while the third is for the unbounded square loss.

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

Choosing as prior an *empirical Gibbs* distribution $Q_s^0(dh) \propto e^{-\gamma \hat{L}(h,s)}\mu(dh)$ for some fixed $\gamma > 0$ and base measure $\mu$ over $\mathcal{H}$, we derive a novel PAC-Bayes bound. Recall that $s$ is the size-$n$ sample. We use $F(x,y) = \sqrt{n}(x-y)$, 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 kernel $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over size-$n$ i.i.d. samples $S$ we have

$$Q_S[L] - Q_S[\hat{L}_S] \leq \frac{1}{\sqrt{n}}\left(\mathrm{KL}(Q_S\|Q_S^0) + 2\Big(1+\frac{2\gamma}{\sqrt{n}}\Big) + \log\Big(\frac{1+\sqrt{e}}{\delta}\Big)\right) . \tag{6}$$

Notice that this prior allowed to remove '$\log(n)$' from the usual PAC-Bayes bounds (see our Eq. (1) and Eq. (2) above). This was one of the important contributions of Catoni [2007], who also used a data-dependent Gibbs distribution, see Catoni [2007, Theorem 1.2.4, Theorem 1.3.1, & corollaries]. 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: $Q_S[L] - Q_S[\hat{L}_S] \lesssim 1/\sqrt{n} + \gamma/n$ . Notice that this latter bound has 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]. Essentially, they show (see 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[9], one can turn the inequality $F(Q_S[L], Q_S[\hat{L}_S]) \leq \mathrm{KL}(Q_S\|Q^*) + \log(\xi(Q^*)/\delta)$ into

$$F(Q_S[L], Q_S[\hat{L}_S]) \leq \mathrm{KL}(Q_S\|Q_S^0) + \log(2\xi(Q^*)/\delta) + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\log(\frac{4}{\delta})} . \tag{7}$$

In other words, if Eq. (4) holds with a data-free prior $Q^*$, then Eq. (7) holds with a data-dependent prior that is distributionally stable (i.e. satisfies $\mathrm{DP}(\epsilon)$). Note that different choices of $F$ would lead to different bounds on $\xi(Q^*)$ —essentially, upper bounds on the exponential moment typically considered in the PAC-Bayesian literature. For example, taking $F(x,y) = n\,\mathrm{kl}(y\|x)$ one can show that $\xi(Q^*) \leq 2\sqrt{n}$ [Maurer, 2004], and this leads to Theorem 4.2 of Dziugaite and Roy [2018b]: if $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{H})$ satisfies the $\mathrm{DP}(\epsilon)$ property, then for any kernel $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over size-$n$ i.i.d. samples $S$ we have

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

Eq. (7) is a general version of this result, whose derivation is based on the notion of *max-information* [Dwork et al., 2015a]. The details of the general conversion recipe are given in Appendix C.

[9] $\mathrm{DP}(\epsilon)$ stands for "differential privacy with $\epsilon$." See Appendix C for details on this property.

### 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 unbounded squared loss and a data-dependent prior. 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 $h_w : \mathbb{R}^d \to \mathbb{R}$ 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 $\mathcal{W} = \mathbb{R}^d$. The size-$n$ random sample is $S = ((X_1, Y_1), \dots, (X_n, Y_n)) \in (\mathbb{R}^d \times \mathbb{R})^n$. The population and empirical losses are defined with respect to the square loss function:

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

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

Consider the prior $Q^0_{\gamma,\lambda}$ with density $q^0_{\gamma,\lambda}(w) \propto e^{-\frac{\gamma\lambda}{2}\|w\|^2}$ 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 $\lambda > \max_i\{\lambda_i - \hat{\lambda}_i\}$, with probability one over size-$n$ random samples $S$ we have

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

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.

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 KL(Posterior$\|$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 $|\hat{\lambda}_i - \lambda_i| \lesssim \sqrt{d/n}$ and the last term in Eq. (8) then behaves as $d\log\big(\lambda/(\lambda + \hat{\lambda}_i - \lambda_i)\big) \lesssim d/\sqrt{n-1}$. 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].

The explicit dependence on the spectrum of the sample covariance matrix opens interesting venues for distribution-dependent analysis. The above argument can be extended to heavy-tailed data distributions, where in some cases we can have concentration of the smallest eigenvalue of a sample covariance matrix even for unbounded instances, see Vershynin [2011, Section 5.4.2]. Moreover, our technique allows to combine PAC-Bayes analysis with specific applications by considering various data distributions. For instance, we can obtain bounds for structured data by analyzing eigenvalues of the corresponding (sparse or blocked) covariance matrices [Wainwright, 2019], thus revealing fined-grained dependence on the distribution compared to the usual PAC-Bayes bounds. Similarly, one can obtain generalization bounds for statistically dependent data by looking at the concentration of the covariance with dependent observations [de la Peña and Giné, 2012].

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 $\lambda > \max_i\{\lambda_i - \hat{\lambda}_i\}$, with probability one over random samples $S$,

$$\log Q^0_{\gamma,\lambda}[e^f] = \gamma \min_{w\in\mathbb{R}^d}\left(L(w) - (\hat{L}_S(w) + \frac{\lambda}{2}\|w\|^2)\right) + \frac{1}{2}\sum_{i=1}^{d}\log\left(\frac{\lambda}{\lambda + \hat{\lambda}_i - \lambda_i}\right) . \tag{9}$$

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.

Finally, note that Eq. (9) elucidates an equivalence between the concentration of eigenvalues of the sample covariance matrix and concentration of the empirical loss. Indeed, for simplicity assuming a noise-free setting (that is $\min_{w\in\mathbb{R}^d} L(w) = 0$), we observe that whenever $(\hat{\lambda}_i - \lambda_i) \to 0$ as $n \to \infty$ for i.i.d. instances, we have $\hat{L}_S(w) \to L(w)$. This provides an alternative way to control the concentration, compared to works based on restrictions on the loss as e.g. by Germain et al. [2016], Holland [2019]. We discuss another PAC-Bayes bound for unbounded losses in Appendix E.

## 3 Our PAC-Bayes theorem for stochastic kernels

The following results involve data- and hypothesis-dependent functions $f : \mathcal{S}\times\mathcal{H} \to \mathbb{R}$. Notice that the order $\mathcal{S}\times\mathcal{H}$ is immaterial—functions $\mathcal{H}\times\mathcal{S} \to \mathbb{R}$ 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, $\rho[f_s] = \int_{\mathcal{H}} f_s(h)\rho(dh)$. When $\rho$ is data-dependent, that is, $\rho \in \mathcal{K}(\mathcal{S},\mathcal{H})$ 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 $\rho_s[f_s] = \int_{\mathcal{H}} f_s(h)\rho_s(dh)$.

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[10] by $P \otimes Q$ that acts on functions $\phi : \mathcal{S}\times\mathcal{H} \to \mathbb{R}$ as follows:

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

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 $\mathbb{E}[\phi(S,H)] = \mathbb{E}[\mathbb{E}[\phi(S,H)|S]]$. Our basic result is the following theorem.

**Theorem 1 (Basic PAC-Bayes inequality)** *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*

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

*(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*

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

To the best of our knowledge, this theorem is new. Notice that $Q^0$ is by default a stochastic kernel from $\mathcal{S}$ to $\mathcal{H}$. Hence, given data $S$, the prior $Q^0_S$ is a data-dependent distribution over hypotheses. By contrast, the usual PAC-Bayes approaches assume that $Q^0$ is a 'data-free' distribution. Also note that the function $f$ is unrestricted, and the distribution $P \in \mathcal{M}_1(\mathcal{S})$ is unrestricted, except for integrability conditions to ensure that $\xi$ is finite. 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:

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

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

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) Recall $f_S(H) = f(S,H)$. Notice that $\mathbb{E}[D|S] = Q_S[f_S] - \mathrm{KL}(Q_S\|Q^0_S)$ . By Jensen inequality, $\mathbb{E}[D|S] \leq \log \mathbb{E}[e^D|S]$. 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 $\log \mathbb{E}[e^D|S] \leq \log \mathbb{E}[e^D] + \log(1/\delta)$. ■

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 $F(Q_s[A_s]) \leq Q_s[F(A_s)]$ 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 $A : \mathcal{S} \times \mathcal{H} \to \mathbb{R}^k$ 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 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*

$$F(Q_S[A_S]) \leq \mathrm{KL}(Q_S\|Q^0_S) + \log(\xi/\delta) . \tag{10}$$

This theorem is a general template for deriving PAC-Bayes bounds, not just with 'data-free' priors, but also more generally with data-dependent priors. Previous works (see Section 4 below) that presented similar generic templates for deriving PAC-Bayes bounds only considered data-free priors. We emphasize that a 'data-free' distribution is equivalent to a constant stochastic kernel: $Q^0_s = Q^0_{s'}$ for all $s, s' \in \mathcal{S}$. Hence $\mathcal{M}_1(\mathcal{H}) \subset \mathcal{K}(\mathcal{S},\mathcal{H})$, which implies that our Theorem 2 encompasses the usual PAC-Bayes inequalities with data-free priors in the literature.

Interestingly, our Theorem 2 is valid with any normed space instead of $\mathbb{R}^k$. This theorem extends the typically used case where $k = 2$ and $A = (L(h), \hat{L}(h,s))$, in which case the function of interest is $f(s,h) = F(L(h), \hat{L}(h,s))$, where $F : \mathbb{R}^2 \to \mathbb{R}$ is a convex function, but there are no restrictions on the loss function $\ell$ that is used in defining $L(h)$ and $\hat{L}(h,s)$. Hence Theorem 2 is valid for *any* loss function: convex or non-convex, bounded or unbounded. Notice also that our Theorem 2 holds for any $P \in \mathcal{M}_1(\mathcal{S})$, i.e. without restrictions on the data-generating process. In particular, our Theorem 2 holds without the i.i.d. data assumption, hence this theorem could potentially enable new generalization bounds for statistically dependent data. In Section 4 below we comment on some literature related to unbounded losses and non-i.i.d. data.

An important role is played by $\xi$, the exponential moment (moment generating function at 1) of the function $f$ under the joint distribution $P \otimes Q^0$. As discussed above in Section 2, there are essentially two main steps involved in obtaining a PAC-Bayesian inequality: (i) choose $F$ to use in Theorem 2, and (ii) upper-bound the exponential moment $\xi$. We emphasize that the "usual assumptions" on which PAC-Bayes bounds are based, namely, (a) data-free prior, (b) bounded loss, and (c) i.i.d. data, played a role only in the technique used for controlling $\xi$. This is because with a data-free $Q^0$ we may swap the order of integration:

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

Then bounding $\xi$ proceeds by calculating or bounding $\xi_{\mathrm{swap}}$ for which there are readily available techniques for bounded loss functions and i.i.d. data (see e.g. Maurer [2004], Germain et al. [2009], van Erven [2014]). The bounds with data-dependent priors that we presented in Section 2 required different kinds of techniques to control the exponential moment, the details are in the appendices. To the best of our knowledge, ours is the first work to extend the PAC-Bayes analysis to stochastic kernels. This framework appears to be a promising theoretical tool to obtain new results. The three types of data-dependant priors discussed in Section 2 show the versatility of the approach. Deriving more cases of PAC-Bayes inequalities without the usual assumptions is left for future research.

## 4 Additional discussion and related literature

The literature on the PAC-Bayes learning approach is vast. We briefly mention the usual references McAllester [1999], Langford and Seeger [2001], Seeger [2002], and Catoni [2007]; but see also Maurer [2004], and Keshet et al. [2011]. Note that McAllester [1999] continued McAllester [1998] whose work was inspired by Shawe-Taylor and Williamson [1997]'s work on a PAC analysis of a Bayesian-style estimator. We acknowledge the tutorials of Langford [2005] and McAllester [2013], the mini-tutorial of van Erven [2014], and the primer of Guedj [2019]. Our Theorem 2 is akin to general forms of the PAC-Bayes theorem given before by Audibert [2004], Germain et al. [2009], and Bégin et al. [2014, 2016]. Our Theorem 1(i) is akin to the "pointwise" bound of Blanchard and Fleuret [2007], in that the bound holds over the random draw of data and hypothesis pairs.

There are many application areas that have used the PAC-Bayes approach, but there are essentially two ways that a PAC-Bayes bound is typically applied: either use the bound to give a risk certificate for a randomized predictor learned by some method, or turn the bound itself into a learning method by searching a randomized predictor that minimizes the bound. The latter is mentioned already by 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], possibility among others. Recently, the use of the latter approach has also found success in training neural networks, see 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], Pérez-Ortiz et al. [2020]) 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.

A line of work related to connecting PAC-Bayes 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 the 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. Data-dependent priors were investigated recently by Awasthi et al. [2020], who relied on tools from the empirical process theory and controlled the capacity of a data-dependent hypothesis class (see also Foster et al. [2019]). The PAC-Bayes literature does contain a line of work that investigates relaxing the restriction of bounded loss functions. A straightforward way to extend PAC-Bayes inequalities to unbounded loss functions 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. Recent work has also looked into the analysis for heavy-tailed losses. For example, Alquier and Guedj [2018] proposed a polynomial moment-dependent bound with $f$-divergence replacing the KL divergence, while Holland [2019] devised an exponential bound assuming that the second moment of the loss is bounded uniformly across hypotheses. An alternative approach was explored by Kuzborskij and Szepesvári [2019], who proposed a stability-based approach by controlling the Efron-Stein variance proxy of the loss. Squared loss regression was studied by Shalaeva et al. [2020] who improved results of Germain et al. [2016] and also relaxed the data-generation assumption to non-iid data. It is worth mentioning the important work related to extending the PAC-Bayes framework to statistically dependent data, see e.g. Alquier and Wintenberger [2012] who applied Rio [2000]'s version of Hoeffding's inequality, derived PAC-Bayes bounds for non-i.i.d. data, and used them in model selection for time series.

As we mentioned in the introduction, besides randomized predictions, other prediction schemes may be derived from a learned distribution over hypotheses. Aggregation by exponential weighting was considered by Dalalyan and Tsybakov [2007, 2008], ensembles of decision trees were considered by Lorenzen et al. [2019], weighted majority vote by Masegosa et al. [2020], Germain et al. [2015]. This list is far from being complete. 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, $Q_S[L] - \inf_{h\in\mathcal{H}} L(h)$, we refer e.g. to Catoni [2007], Alquier et al. [2016], Grünwald and Mehta [2019], Kuzborskij et al. [2019], Mhammedi et al. [2019]. The approach of analyzing the gap (for randomized predictors), which we follow in this paper, is generally complementary to such excess risk analyses.

## Broader Impact

We think this work will have a positive impact on the theoretical machine learning community. However, since this work presents a high-level theoretical framework, its direct impact on society will be linked to the particular user-specific applications where this framework may be instantiated.

## Acknowledgments and Disclosure of Funding

We warmly thank the anonymous reviewers for their valuable feedback, which helped us to improve the paper greatly. For comments on various early parts of this work we warmly thank Tor Lattimore, Yevgeny Seldin, Tim van Erven, Benjamin Guedj, and Pascal Germain. We warmly acknowledge the Foundations team at Deepmind, and the AI Centre at University College London, for providing friendly and stimulating work environments. Omar Rivasplata and Ilja Kuzborskij warmly thank Vitaly Feldman for interesting discussions and a fun table tennis game while visiting DeepMind.

Omar Rivasplata gratefully acknowledges DeepMind sponsorship for carrying out research studies on the theoretical foundations of machine learning and AI at University College London. This work was done while Omar was a research scientist intern at DeepMind.

Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, the Alberta Machine Intelligence Institute (Amii), and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

John Shawe-Taylor gratefully acknowledges support and funding from the U.S. Army Research Laboratory and the U. S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1.

## References

P Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

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

P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

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 *Advances 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.

P. Awasthi, S. Kale, S. Karp, and M. Mohri. PAC-Bayes Learning Bounds for Sample-Dependent Priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian theory for transductive learning. In *Artificial Intelligence and Statistics (AISTATS)*, 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 (AISTATS)*, 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.

G. Blanchard and F. Fleuret. Occam's hammer. In *Conference on Learning Theory (COLT)*, pages 112–126. Springer, 2007.

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

O. Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer, 2004.

O. Catoni. PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. IMS Lecture Notes-Monograph Series, 56, 2007. URL www.jstor.org/stable/20461499.

T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, 2nd. edition, 2006.

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

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Conference on Learning Theory (COLT)*, pages 97–111. Springer, 2007.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

V. de la Peña and E. Giné. *Decoupling: from dependence to independence*. Springer, 2012.

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. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2350–2358, 2015a. Our citations refer to the full version arXiv:1506.02629.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8430–8441, 2018b.

S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. Wiley, 1986.

D. J. Foster, S. Greenberg, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Hypothesis Set Stability and Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6729–6739, 2019.

Y. Freund. Self bounding learning algorithms. In *Conference on Learning Theory (COLT)*, 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, A. Lacasse, F. Laviolette, M. Marchand, and J.-F. Roy. Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm. *Journal of Machine Learning Research*, 16:787–860, 2015.

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances 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 *Algorithmic Learning Theory (ALT)*, volume 98, pages 433–465. PMLR, 2019.

B. Guedj. A Primer on PAC-Bayesian Learning. arXiv:1901.05353, 2019.

M. Holland. PAC-Bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2715–2724, 2019.

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 *Conference on Learning Theory (COLT)*, volume 99, pages 2028–2054. PMLR, 2019.

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 *Advances 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 *Advances in Neural Information Processing Systems (NIPS)*, pages 2931–2940, 2017.

S. S. Lorenzen, C. Igel, and Y. Seldin. On PAC-Bayesian bounds for random forests. *Machine Learning*, 108(8-9):1503–1522, 2019.

A. R. Masegosa, S. S. Lorenzen, C. Igel, and Y. Seldin. Second Order PAC-Bayesian Bounds for the Weighted Majority Vote. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2007.13532.

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

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.

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

F. McSherry and K. Talwar. Mechanism Design via Differential Privacy. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, volume 7, pages 94–103. IEEE, 2007.

S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd. edition, 2009.

Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes Un-Expected Bernstein Inequality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12202–12213, 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 (AISTATS)*, 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. Pérez-Ortiz, O. Rivasplata, J. Shawe-Taylor, and C. Szepesvári. Tighter risk certificates for neural networks. arXiv:2007.12911, 2020.

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.

E. Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 330(10):905–908, 2000.

O. Rivasplata, E. Parrado-Hernández, J. Shawe-Taylor, S. Sun, and C. Szepesvári. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In *Advances 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 generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3(Oct):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.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

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. PAC-Bayesian ensemble learning. Master's thesis, University of Copenhagen, 2016.

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 *Advances 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:1011.3027, 2011. Chapter 5 of: Compressed Sensing, Theory and Applications. Edited by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012. pp. 210–268.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

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

# PAC-Bayes Analysis Beyond the Usual Bounds: Supplementary Material

**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

## 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 $\nu$-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 $P \in \mathcal{M}_1(\mathcal{X})$, i.e. for sets $A \in \Sigma_{\mathcal{X}}$ the event that the value of $X$ falls within $A$ has probability $\mathbb{P}[X \in A] = P(A)$. Then the expectation of $f(X)$ is $\mathbb{E}[f(X)] = P[f]$, and its variance is $\mathrm{Var}[f(X)] = P[f^2] - P[f]^2$.

## 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 $\mathcal{S} \times \mathcal{H}$ 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 $\mathbb{E}[\phi(S, H)] = \mathbb{E}[\mathbb{E}[\phi(S, H)|S]]$. Also recall $\mathcal{S} = \mathcal{Z}^n$.

**Lemma 3** *For any $n$, for any loss function with range $[0, b]$, for any $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ such that $Q_s(dh) \propto e^{-\gamma \hat{L}(h,s)}\mu(dh)$, the following upper bound on $\xi(Q) = \mathbb{E}[e^{\sqrt{n}\left(L(H) - \hat{L}(H,S)\right)}]$ 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 $\Delta_s(h) = \sqrt{n}\big(L(h) - \hat{L}(h, s)\big)$ where $(s, h) \in \mathcal{S} \times \mathcal{H}$. 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*

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

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

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

**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*

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

*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*

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

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 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:

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

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

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

so then rearranging

$$\begin{aligned} \log \mathbb{E}\left[e^{\Delta_S(H)}\right] &\leq \mathbb{E}\left[\frac{e^{\Delta_S(H)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]} \left(\Delta_S(H) - (\Delta_S(H'))_+\right)\right] + \log \mathbb{E}\left[e^{(\Delta_S(H'))_+}\right] \\ &\leq \mathbb{E}\left[\frac{e^{\Delta_S(H)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]} \Delta_S(H)\right] + \log \mathbb{E}\left[e^{(\Delta_S(H'))_+}\right] . \end{aligned} \tag{11}$$

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

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

Then the inequality (11) can be written as

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

**Bounding $(I)$.** We handle the first term through the stability analysis of the density $\mu_S$. We will denote by $S^{(i)} = (Z_{1:i-1}, Z'_1, Z_{i+1:n})$ 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,

$$\begin{aligned}\frac{1}{\sqrt{n}}\,\mathbb{E}\int \Delta_S(h)\,\mathrm{d}\mu_S(h) &= \mathbb{E}\int \ell(h,Z'_1)\,\mathrm{d}\mu_S(h) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\int \ell(h,Z_i)\,\mathrm{d}\mu_S(h)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\int \left(\ell(h,Z'_1)-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h) \qquad (12)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\int \ell(h,Z_i)\,\mathrm{d}\mu_{S^{(i)}}(h) - \int \ell(h,Z_i)\,\mathrm{d}\mu_S(h)\right].\end{aligned}$$

The last equality comes from switching $Z'_1$ and $Z_i$ since these variables are distributed identically. Now we use Lemma 4 with $\mu_{S^{(i)}}$ and $\mu_S$, 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 4 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 $q_S$, $\mu_S$, $\mu_{S^{(i)}}$ are absolutely continuous,

$$\begin{aligned}\mathrm{KL}\left(\mu_{S^{(i)}}\|\mu_S\right) &= \int \log(\mathrm{d}\mu_{S^{(i)}}(h)/\,\mathrm{d}h)\,\mathrm{d}\mu_{S^{(i)}}(h) - \int \log(\mathrm{d}\mu_S(h)/\,\mathrm{d}h)\,\mathrm{d}\mu_{S^{(i)}}(h)\\ &= \int \log\left(\frac{e^{\Delta_{S^{(i)}}(h)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]}\frac{e^{-\gamma\hat{L}_{S^{(i)}}(h)}}{N_{S^{(i)}}}\right)\mathrm{d}\mu_{S^{(i)}}(h) - \int \log\left(\frac{e^{\Delta_S(h)}}{\mathbb{E}\left[e^{\Delta_S(H)}\right]}\frac{e^{-\gamma\hat{L}_S(h)}}{N_S}\right)\mathrm{d}\mu_{S^{(i)}}(h)\\ &= \int \left(\Delta_{S^{(i)}}(h)-\Delta_S(h)\right)\mathrm{d}\mu_{S^{(i)}}(h) + \log\left(\frac{N_S}{N_{S^{(i)}}}\right) + \gamma\int\left(\hat{L}_S(h)-\hat{L}_{S^{(i)}}(h)\right)\mathrm{d}\mu_{S^{(i)}}(h)\\ &\le \sqrt{n}\int\left(\hat{L}_S(h)-\hat{L}_{S^{(i)}}(h)\right)\mathrm{d}\mu_{S^{(i)}}(h) \qquad \text{(By definition of } \Delta_S)\\ &\quad + \gamma\int\left(\hat{L}_{S^{(i)}}(h)-\hat{L}_S(h)\right)\mathrm{d}\mu_S(h) \qquad \text{(By Lemma 5)}\\ &\quad + \gamma\int\left(\hat{L}_S(h)-\hat{L}_{S^{(i)}}(h)\right)\mathrm{d}\mu_{S^{(i)}}(h)\\ &= \frac{1}{\sqrt{n}}\int\left(\ell(h,Z_i)-\ell(h,Z'_1)\right)\mathrm{d}\mu_{S^{(i)}}(h)\\ &\quad + \frac{\gamma}{n}\int\left(\ell(h,Z'_1)-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h)\\ &\quad + \frac{\gamma}{n}\int\left(\ell(h,Z_i)-\ell(h,Z'_1)\right)\mathrm{d}\mu_{S^{(i)}}(h)\,,\end{aligned}$$

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. (12) (each $i \in [n]$) we get

$$\begin{aligned}&\mathbb{E}\left[\int\left(\ell(h,Z'_1)-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h)\right] = \mathbb{E}\left[\int \ell(h,Z_i)\,\mathrm{d}\mu_{S^{(i)}}(h) - \int \ell(h,Z_i)\,\mathrm{d}\mu_S(h)\right]\\ &\le \mathbb{E}\left[\sqrt{2b^2\,\mathrm{KL}\left(\mu_{S^{(i)}}\|\mu_S\right)}\right] \le \sqrt{2b^2\,\mathbb{E}[\mathrm{KL}\left(\mu_{S^{(i)}}\|\mu_S\right)]} \qquad \text{(By Lemma 4 and Jensen)}\\ &= \sqrt{2b^2\left(\frac{1}{\sqrt{n}}+\frac{2\gamma}{n}\right)\mathbb{E}\left[\int\left(\ell(h,Z'_1)-\ell(h,Z_i)\right)\mathrm{d}\mu_S(h)\right]}\,.\end{aligned}$$

The last calculation implies

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

Finally, combining this with Eq. (12) gives

$$\mathbb{E}\int \Delta_S(h)\,\mathrm{d}\mu_S(h) \leq 2b^2\left(1+\frac{2\gamma}{\sqrt{n}}\right). \tag{13}$$

**Bounding $(II)$.** Now we turn our attention to the exponential moment of $(\Delta_S(H'))_+$ in (11):

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

and observe that the internal expectation is bounded as

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

where we obtain the last inequality thanks to Hoeffding's lemma for independent random variables with values in the range $[-b/\sqrt{n}, b/\sqrt{n}]$. Plugging the bounds on terms $(I)$ and $(II)$ into Eq. (11) finishes the proof of Lemma 3. ■

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

**Corollary 6** *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 $Q^0_s \propto e^{-\gamma\hat{L}(h,s)}$, for any $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over size-$n$ i.i.d. samples $S\sim P_1^n$ we have*

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

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

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

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

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

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 $\pi \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ be a stochastic kernel. Recall that $\mathcal{S} = \mathcal{Z}^n$ is the space of size-$n$ samples. When we say that $\pi$ 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 condition on the Radon-Nikodym derivative is equivalent to the condition that, whenever $s$ and $s'$ differ at one entry, the ratio $\pi(s, A)/\pi(s', A)$ is upper bounded by $e^{\epsilon}$, for all sets $A \in \Sigma_{\mathcal{H}}$. Thus, the property entails stability of the data-dependent distribution $\pi_s$ with respect to small changes in the composition of the $n$-tuple $s$. This definition goes back to the literature on privacy-preserving methods for data analysis [Dwork et al., 2015b]; however, we are interested in its formal properties only. It captures a kind of 'distributional stability' which we refer to as 'd-stability' for short.

As noted before, the main challenge in obtaining PAC-Bayes bounds is in controlling the exponential moment $\xi(Q^0) = (P_n \otimes Q^0)[e^f]$ for given $P_n \in \mathcal{M}_1(\mathcal{S})$ and $Q^0 \in \mathcal{K}(\mathcal{S}, \mathcal{H})$.

In the following we rely on a notion of $\beta$-approximate *max-information* [Dwork et al., 2015a,b], denoted $I_\infty^\beta(X;Y)$ for $\beta > 0$ and arbitrary random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. Intuitively, this intends to measure the worst-case 'distributional distance' of the jointly distributed pair $(X, Y)$ from the pair $(X', Y)$ with $X'$ a copy of $X$ independent from $Y$. Formally, $I_\infty^\beta(X;Y)$ is defined as the least $\eta > 0$ such that for every $C \in \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}}$ (the product sigma-algebra) we have

$$\mathbb{P}[(X, Y) \in C] \leq e^{\eta}\mathbb{P}[(X', Y) \in C] + \beta .$$

Special care is needed in defining $I_\infty^\beta(X; \psi(X))$, i.e. the 'distributional distance' of the pair $(X, \psi(X))$ to the independent pair $(X', \psi(X))$. In our context (see below) we need $I_\infty^\beta(S; Q_S^0)$. The next lemma generalizes an idea we learned from Dziugaite and Roy [2018b]:

**Lemma 7** (max-information lemma) *Fix $n \in \mathbb{N}$, $P_n \in \mathcal{M}_1(\mathcal{S})$, and a function $f : \mathcal{S} \times \mathcal{H} \to \mathbb{R}$. Let $\zeta(n)$ be a positive sequence (possibly constant). Suppose that for any data-free distribution $Q^* \in \mathcal{M}_1(\mathcal{H})$, for any kernel $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$ and for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over size-$n$ random samples $S \sim P_n$ the following holds:*

$$Q_S[f_S] \leq \mathrm{KL}(Q_S \| Q^*) + \log(\zeta(n)/\delta) . \tag{14}$$

*Then for any kernels $Q^0, Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$, and for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over size-$n$ random samples $S \sim P_n$ we have*

$$Q_S[f_S] \leq \mathrm{KL}(Q_S \| Q_S^0) + \log(2\zeta(n)/\delta) + I_\infty^{\alpha/2}(S; Q_S^0) . \tag{15}$$

This lemma gives a general recipe for converting a PAC-Bayes bound with a fixed 'data-free' prior (i.e. Eq. (14)) into a similar PAC-Bayes bound with a data-dependent prior (Eq. (15)). The choice of $\zeta(n)$ is problem-dependent, but the idea is that if $\xi(Q^*) = (P_n \otimes Q^*)[e^f]$ satisfies $\xi(Q^*) \leq \zeta(n)$ when $Q^*$ is a data-free distribution, then $\zeta(n)$ can be re-used in Eq. (15). For a given $P_n$ and $f$, the best choice of $\zeta(n)$ would be $\zeta(n) = \inf_{Q^* \in \mathcal{M}_1(\mathcal{H})} \int \int e^{f(s,h)} Q^*(dh) P_n(ds)$.

The statement of Lemma 7 is written in the generic framework of Theorem 1. We may specialize it to Theorem 2 when the function $f$ used in the left hand side of the inequality—and in the exponential moment $\xi(Q^*) = (P_n \otimes Q^*)[e^f]$—has the form of a composition $f(s, h) = F(A(s, h))$, with $A : \mathcal{S} \times \mathcal{H} \to \mathbb{R}^k$ any measurable function, and $F : \mathbb{R}^k \to \mathbb{R}$ any convex function. The literature uses $k = 2$ and $A = (L(h), \hat{L}(h, s))$; and various choices of $F$ lead to various PAC-Bayes bounds. Notice that, by Jensen's inequality, $F(Q_s[A_s]) \leq Q_s[F(A_s)] = Q_s[f_s]$ for any $s$.

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 stochastic kernel $Q^0$ satisfies the $\mathrm{DP}(\epsilon)$ property:

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

Therefore, via the max-information lemma, one may derive PAC-Bayes bounds which are valid for d-stable data-dependent priors. Specific forms of the upper bound can be obtained when a specific $\zeta(n)$ (i.e. a bound on $\xi(Q^*)$) is available. For instance, for the PAC-Bayes-kl bound, which uses $F(x, y) = n\,\mathrm{kl}(y\|x)$, we may take $\zeta(n) = 2\sqrt{n}$ [Maurer, 2004], and obtain the following:

**Theorem 8** *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 at least $1-\delta$ over size-$n$ i.i.d. samples $S \sim P_1^n$ we have*

$$\mathrm{kl}(Q_S[\hat{L}_S] \| Q_S[L]) \leq \frac{\mathrm{KL}(Q_S \| Q_S^0) + \log(\frac{4\sqrt{n}}{\delta}) + \frac{n\epsilon^2}{2} + \epsilon\sqrt{\frac{n}{2}\log(\frac{4}{\delta})}}{n} . \tag{16}$$

This is Theorem 4.2 of Dziugaite and Roy [2018b]. The proof of this theorem takes as starting point the PAC-Bayes-kl bound [Seeger, 2002, Langford and Seeger, 2001], which says that when $Q^* \in \mathcal{M}_1(\mathcal{H})$ is a data-free distribution over hypotheses, for any $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over size-$n$ i.i.d. samples $S \sim P_1^n$ we have

$$\mathrm{kl}(Q_S[\hat{L}_S] \| Q_S[L]) \leq \frac{\mathrm{KL}(Q_S \| Q_S^0) + \log(\xi(Q^*)/\delta)}{n} .$$

Notice that this PAC-Bayes-kl inequality follows from Theorem 2, which in turn follows from Theorem 1(ii), using $f(s,h) = F(L(h), \hat{L}(h,s))$ with $F(x,y) = n\,\mathrm{kl}(y\|x)$ under the restriction of losses within the range $[0,1]$. Then we may use $\xi(Q^*) \leq 2\sqrt{n}$ since $Q^*$ is a fixed 'data-free' distribution (cf. Maurer [2004]). Then use Lemma 7, and upper-bound the $(\alpha/2)$-approximate max-information as per the inequality of Dwork et al. [2015a, Theorem 20] cited before Theorem 8.

### 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. Let $Q^* \in \mathcal{M}_1(\mathcal{H})$ be a fixed 'data-free' distribution over $\mathcal{H}$, and let $Q \in \mathcal{K}(\mathcal{S},\mathcal{H})$ be a stochastic kernel. Suppose Eq. (14) is satisfied (this is the assumption required by Lemma 7). Given $\delta' \in (0,1)$, define the set

$$\mathcal{E}(Q^*) = \left\{ s \in \mathcal{S} \;\middle|\; Q_s[f_s] > \mathrm{KL}(Q_s \| Q^*) + \log(\zeta(n)/\delta') \right\} .$$

Notice that for a random sample $S \sim P_n$ we have $\mathbb{P}[S \in \mathcal{E}(Q^*)] = P_n(\mathcal{E}(Q^*)) \leq \delta'$ by Eq. (14). Now suppose $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{H})$ is a stochastic kernel, so each random size-$n$ data set $S$ is mapped to a data-dependent distribution $Q_S^0$ over $\mathcal{H}$. Correspondingly, define the set

$$\mathcal{E}(Q^0) = \left\{ (s,s') \in \mathcal{S}\times\mathcal{S} \;\middle|\; Q_s[f_s] > \mathrm{KL}(Q_s \| Q_{s'}^0) + \log(\zeta(n)/\delta') \right\} .$$

We are interested in the event that a random sample $S \sim P_n$ satisfies $(S,S) \in \mathcal{E}(Q^0)$. For fixed $s' \in \mathcal{S}$, consider the section $\mathcal{E}(Q^0)_{s'} = \{s \in \mathcal{S} \mid (s,s') \in \mathcal{E}(Q^0)\}$; and notice that $(s,s') \in \mathcal{E}(Q^0)$ if and only if $s \in \mathcal{E}(Q^0)_{s'}$. For any fixed $s'$, the random sample satisfies $\mathbb{P}[S \in \mathcal{E}(Q^0)_{s'}] \leq \delta'$, again by Eq. (14). Then if $S' \sim P_n$ is an independent copy of $S$, we have

$$\mathbb{P}[(S,S') \in \mathcal{E}(Q^0)] = \mathbb{P}[S \in \mathcal{E}(Q^0)_{S'}] = \mathbb{E}[\mathbb{P}[S \in \mathcal{E}(Q^0)_{S'} | S']] \leq \delta' .$$

By the definition of $\beta$-approximate max-information [Dwork et al., 2015a] we have

$$\mathbb{P}[(S,S) \in \mathcal{E}(Q^0)] \leq e^{I_\infty^\beta(S;Q_S^0)} \mathbb{P}[(S,S') \in \mathcal{E}(Q^0)] + \beta \leq e^{I_\infty^\beta(S;Q_S^0)}\delta' + \beta .$$

Therefore, given $\delta \in (0,1)$, setting $\beta = \delta/2$ and $\delta' = e^{-I_\infty^{\alpha/2}(S;Q_S^0)}\delta/2$, we get $\mathbb{P}[S \in \mathcal{E}(Q^0)_S] \leq \delta$. This finishes the proof of the "max-information lemma" (Lemma 7).

**Remark.** Let $Q^* \in \mathcal{M}_1(\mathcal{H})$ be a 'data-free' distribution, and suppose the exponential moment $\xi(Q^*) = \int\int e^{f(s,h)} Q^*(dh) P_n(ds)$ satisfies $\xi(Q^*) \leq \xi_{\mathrm{bd}}$. If a stochastic kernel $Q^0 \in \mathcal{K}(\mathcal{S},\mathcal{H})$ satisfies $\mathrm{DP}(\epsilon)$ for some $\epsilon > 0$, then in the exponential moment

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

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

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

where the integral on the right hand side is upper bounded by $\xi_{\mathrm{bd}}$ 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 so that $\log(\xi(Q^0))$ is 'replaced' with $\log(2\xi_{\mathrm{bd}}) + I_\infty^{\delta/2}(S;Q_S^0)$; whereas the naive argument just described would give $\log(\xi(Q^0)) \leq \log(\xi_{\mathrm{bd}}) + n\epsilon$.

## 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 $h_w : \mathbb{R}^d \to \mathbb{R}$ 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 $\mathcal{W} = \mathbb{R}^d$. 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 $\Delta_w^S = L(w) - \hat{L}_S(w)$, defined for $w \in \mathbb{R}^d$, where

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

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

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

$$\int_{\mathbb{R}^d} \Delta_w^S\, q_S(w)\,\mathrm{d}w \leq \mathrm{KL}(q_S \,||\, q^0) + \log \int_{\mathbb{R}^d} e^{\Delta_w^S}\, q^0(w)\,\mathrm{d}w\,. \tag{17}$$

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. $P(A) = \int_A p(w)\,\mathrm{d}w$ 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 $\mathcal{S}$ to $\mathbb{R}^d$, for any $\gamma > 0$ and $\lambda > \max_i\{\lambda_i - \hat{\lambda}_i\}$, with probability one over random samples $S$,*

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

**Proof** We get the statement by combining Eq. (17) with the analytic form of exponential moment of $\gamma\Delta_w^S$ given by Lemma 10 below. ∎

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

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

This lemma fills in the main part of the proof of Theorem 9. Notice that this lemma computes explicitly the exponential moment of $\gamma\Delta_w^S$, 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 $\min_{w\in\mathbb{R}^d} \Delta_w^{S,\lambda}$ 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 $w^\star \in \arg\min_{w\in\mathbb{R}^d} L(w)$, so that $L(w^\star) = \min_{w\in\mathbb{R}^d} L(w)$, then with probability one over random samples $S$ we have that*

$$\min_{w\in\mathbb{R}^d} \Delta_w^{S,\lambda} \leq L(w^\star)\,.$$

*If $\max_i (X_i^\top w^\star - Y_i)^2 \leq B$ 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 $q_S(w) \propto e^{-\frac{\gamma}{2}\hat{L}_{S,\alpha}(w)}$ and $q^0(w) \propto e^{-\frac{\gamma\lambda}{2}\|w\|^2}$ and any $\alpha, \lambda, \gamma > 0$,*

$$\mathrm{KL}(q_S \,||\, q^0) = \frac{1}{2}\left(\log\det\left(\frac{1}{\lambda}\hat{\boldsymbol{\Sigma}}_\alpha\right) + \mathrm{tr}\left(\lambda\hat{\boldsymbol{\Sigma}}_\alpha^{-1} - \mathbf{I}\right) + \frac{\lambda\gamma}{n^2}\sum_{i=1}^n Y_i^2 \|X_i\|^2_{\hat{\boldsymbol{\Sigma}}_\alpha^{-2}}\right)\,.$$

*Furthermore, if $\max_i \|X_i\|_2 \leq 1$ a.s., then*

$$\mathrm{KL}(q_S \,||\, q^0) \leq \frac{1}{2}\left(d\log\left(\frac{1+\alpha}{\lambda}\right) + d\left(\frac{\lambda}{\hat{\lambda}_d + \alpha} - 1\right) + \frac{\lambda\gamma}{n^2}\sum_{i=1}^n Y_i^2 \|X_i\|^2_{\hat{\boldsymbol{\Sigma}}_\alpha^{-2}}\right)\,.$$

Combining the results outlined above yields the following corollary.

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

$$\int_{\mathbb{R}^d} \Delta_w^S\, q_S(w)\,\mathrm{d}w \leq \min_{w\in\mathbb{R}^d} \Delta_w^{S,c\hat{\varepsilon}_n} + \frac{d}{2\gamma}\log\left(\frac{1+\alpha}{e(c-1)\hat{\varepsilon}_n}\right) + \frac{c\hat{\varepsilon}_n d}{\hat{\lambda}_d + \alpha}\left(\frac{1}{2\gamma} + \frac{1}{n}\sum_{i=1}^n Y_i^2\right)\,.$$

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 $\gamma^{-1}\,\mathrm{KL}(q_S \,||\, q^0_{\gamma,\lambda}) \to 0$ 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 $\boldsymbol{s} = \mathbb{E}[Y_1 X_1]$ and its empirical counterpart $\hat{\boldsymbol{S}} = (Y_1X_1 + \cdots + Y_nX_n)/n$. Also let's define $C = \mathbb{E}[Y_1^2] - (Y_1^2 + \cdots + Y_n^2)/n$. The density is $q^0(w) = Z_0^{-1} e^{-\frac{\gamma\lambda}{2}\|w\|^2}$, with $Z_0$ a normalizing factor. A straightforward

expression of the integral gives

$$\begin{aligned}
\int_{\mathbb{R}^d} e^{\gamma\left(L(w)-\hat{L}_S(w)\right)} q^0(w)\,\mathrm{d}w &= \frac{1}{Z_0}\int_{\mathbb{R}^d} e^{\gamma\left(L(w)-\hat{L}_{S,\lambda}(w)\right)}\,\mathrm{d}w \\
&= \frac{1}{Z_0}\int_{\mathbb{R}^d} e^{\gamma\left(C-\frac{1}{2}w^\top(\hat{\boldsymbol{\Sigma}}_\lambda-\boldsymbol{\Sigma})w-(\boldsymbol{s}-\hat{\boldsymbol{S}})^\top w\right)}\,\mathrm{d}w \qquad (18)\\
&= \frac{(2\pi)^{\frac{d}{2}}}{Z_0}\frac{e^{\gamma\left(C+\frac{1}{2}(\boldsymbol{s}-\hat{\boldsymbol{S}})^\top(\hat{\boldsymbol{\Sigma}}_\lambda-\boldsymbol{\Sigma})^{-1}(\boldsymbol{s}-\hat{\boldsymbol{S}})\right)}}{\sqrt{\gamma^d\det\left(\hat{\boldsymbol{\Sigma}}_\lambda-\boldsymbol{\Sigma}\right)}} \qquad (19)\\
&= \frac{(2\pi)^{\frac{d}{2}}}{Z_0}\frac{e^{\gamma\min_{w\in\mathbb{R}^d}\left\{L(w)-\hat{L}_{S,\lambda}(w)\right\}}}{\sqrt{\gamma^d\det\left(\hat{\boldsymbol{\Sigma}}_\lambda-\boldsymbol{\Sigma}\right)}} \qquad (20)\\
&= \sqrt{\frac{\lambda^d}{\det\left(\hat{\boldsymbol{\Sigma}}_\lambda-\boldsymbol{\Sigma}\right)}}\,e^{\gamma\min_{w\in\mathbb{R}^d}\left\{L(w)-\hat{L}_{S,\lambda}(w)\right\}} \qquad (21)
\end{aligned}$$

where Eq. (18) is just rewriting things, while in Eq. (19) we assume that $\lambda > \max_i\{\lambda_i - \hat{\lambda}_i\}$. Eqs. (19) and (21) come from Gaussian integration, and Eq. (20) is a consequence of:

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

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

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

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

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

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

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

which completes the proof of Proposition 14. ■

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

$$\Delta_w^{S,\lambda} \le \frac{1}{2}\,\mathbb{E}[(X_1^\top w^\star - Y_1)^2] - \frac{1}{2n}\sum_{i=1}^n (X_i^\top w^\star - Y_i)^2 \le B\sqrt{\frac{x}{2n}}\,.$$

This completes the proof of Proposition 11 ■

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

$$\begin{aligned}
q_S(w) &= \frac{e^{-\frac{\gamma}{2}w^\top\hat{\boldsymbol{\Sigma}}_\alpha w+\gamma w^\top\hat{\boldsymbol{S}}-\frac{\gamma}{2}\bar{Y}^2}}{\int_{\mathbb{R}^d} e^{-\frac{\gamma}{2}u^\top\hat{\boldsymbol{\Sigma}}_\alpha u+\gamma u^\top\hat{\boldsymbol{S}}-\frac{\gamma}{2}\bar{Y}^2}\,\mathrm{d}u} \\
&= \frac{e^{-\frac{\gamma}{2}w^\top\hat{\boldsymbol{\Sigma}}_\alpha w+\gamma w^\top\hat{\boldsymbol{S}}-\frac{\gamma}{2}\hat{\boldsymbol{S}}^\top\hat{\boldsymbol{\Sigma}}_\alpha^{-1}\hat{\boldsymbol{S}}}}{\int_{\mathbb{R}^d} e^{-\frac{\gamma}{2}u^\top\hat{\boldsymbol{\Sigma}}_\alpha u+\gamma u^\top\hat{\boldsymbol{S}}-\frac{\gamma}{2}\hat{\boldsymbol{S}}^\top\hat{\boldsymbol{\Sigma}}_\alpha^{-1}\hat{\boldsymbol{S}}}\,\mathrm{d}u} \\
&=: \mathrm{Gauss}(\hat{\boldsymbol{\Sigma}}_\alpha^{-1}\hat{\boldsymbol{S}},\hat{\boldsymbol{\Sigma}}_\alpha^{-1}) \propto e^{-\frac{\gamma}{2}\left(w-\hat{\boldsymbol{\Sigma}}_\alpha^{-1}\hat{\boldsymbol{S}}\right)^\top\hat{\boldsymbol{\Sigma}}_\alpha\left(w-\hat{\boldsymbol{\Sigma}}_\alpha^{-1}\hat{\boldsymbol{S}}\right)}\,,
\end{aligned}$$

where $\hat{\boldsymbol{S}} = (Y_1 X_1 + \cdots + Y_n X_n)/n$ and $\bar{Y}^2 = (Y_1^2 + \cdots + Y_n^2)/n$. Recall that analytic form of KL-divergence between two Gaussians is:

$$\begin{aligned}
&\mathrm{KL}\left(\mathrm{Gauss}(x_1, \mathbf{A}_1) \,\middle\|\, \mathrm{Gauss}(x_0, \mathbf{A}_0)\right) \\
&\quad = \frac{1}{2}\left(\log\left(\frac{\det \mathbf{A}_0}{\det \mathbf{A}_1}\right) + \mathrm{tr}\left(\mathbf{A}_0^{-1}\mathbf{A}_1\right) - d + (x_1 - x_0)^\top \mathbf{A}_0^{-1}(x_1 - x_0)\right)
\end{aligned}$$

This gives

$$\mathrm{KL}\left(q_S \,||\, q^0\right) = \frac{1}{2}\left(\log\det\left(\frac{1}{\lambda}\hat{\boldsymbol{\Sigma}}_\alpha\right) + \mathrm{tr}\left(\lambda\hat{\boldsymbol{\Sigma}}_\alpha^{-1} - \mathbf{I}\right) + \lambda\gamma\,\hat{\boldsymbol{S}}^\top\hat{\boldsymbol{\Sigma}}_\alpha^{-2}\hat{\boldsymbol{S}}\right)$$

This shows the first statement.

The 'furthermore' statement is shown using a simple fact that for $d \times d$ positive definite matrix $\mathbf{A}$, we have $\det(\mathbf{A}) \le (\mathrm{tr}(\mathbf{A})/d)^d$,

$$\log\det\left(\frac{1}{\lambda}\hat{\boldsymbol{\Sigma}}_\alpha\right) \le d\log\mathrm{tr}\left(\frac{1}{d\lambda}\hat{\boldsymbol{\Sigma}}_\alpha\right) \le d\log\left(\frac{1+\alpha}{\lambda}\right)$$

where we have assumed that $\max_i \|X_i\|_2 \le 1$ a.s. and the fact

$$\mathrm{tr}\left(\lambda\hat{\boldsymbol{\Sigma}}_\alpha^{-1} - \mathbf{I}\right) \le d\left(\frac{\lambda}{\hat{\lambda}_d + \alpha} - 1\right) .$$

This completes the proof of Proposition 12. ■

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

$$\begin{aligned}
\int_{\mathbb{R}^d} \Delta_w^S \, q_S(w)\,\mathrm{d}w &\le \min_{w\in\mathbb{R}^d} \Delta_w^{S,\lambda} + \frac{d}{2\gamma}\log\left(\frac{1+\alpha}{\lambda}\right) + \frac{d}{2\gamma}\left(\frac{\lambda}{\hat{\lambda}_d + \alpha} - 1\right) \\
&\qquad + \frac{\lambda}{2n^2}\sum_{i=1}^n Y_i^2 \|X_i\|^2_{\hat{\boldsymbol{\Sigma}}_\alpha^{-2}} + \frac{1}{2\gamma}\sum_{i=1}^d \log\left(\frac{\lambda}{\lambda + \hat{\lambda}_i - \lambda_i}\right) \\
&\le \min_{w\in\mathbb{R}^d} \Delta_w^{S,c\hat{\varepsilon}_n} + \frac{d}{2\gamma}\log\left(\frac{1+\alpha}{c\hat{\varepsilon}_n}\right) + \frac{d}{2\gamma}\left(\frac{c\hat{\varepsilon}_n}{\hat{\lambda}_d + \alpha} - 1\right) \\
&\qquad + \frac{cd\hat{\varepsilon}_n}{\hat{\lambda}_d + \alpha}\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right) + \frac{d}{2\gamma}\log\left(\frac{c}{c-1}\right) \\
&\le \min_{w\in\mathbb{R}^d} \Delta_w^{S,c\hat{\varepsilon}_n} + \frac{d}{2\gamma}\log\left(\frac{1+\alpha}{e(c-1)\hat{\varepsilon}_n}\right) + \frac{c\hat{\varepsilon}_n d}{\hat{\lambda}_d + \alpha}\left(\frac{1}{2\gamma} + \frac{1}{n}\sum_{i=1}^n Y_i^2\right) ,
\end{aligned}$$

where we used the fact that

$$\sum_{i=1}^d \log\left(\frac{\lambda}{\lambda + \hat{\lambda}_i - \lambda_i}\right) = \sum_{i=1}^d \log\left(\frac{c\max_i\{\lambda_i - \hat{\lambda}_i\}}{c\max_i\{\lambda_i - \hat{\lambda}_i\} - (\lambda_i - \hat{\lambda}_i)}\right) \le d\log\left(\frac{c}{c-1}\right)$$

and by a simple SVD argument

$$\frac{1}{n^2}\sum_{i=1}^n Y_i^2\|X_i\|^2_{\hat{\boldsymbol{\Sigma}}_\alpha^{-2}} \le \frac{d}{n(\hat{\lambda}_d + \alpha)}\sum_{i=1}^n Y_i^2 .$$

This completes the proof of Corollary 13. ■

## E A simple 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 $\lambda > 0$ and $h \in \mathcal{H}$ fixed, we may upper-bound the exponential moment $\mathbb{E}[\exp\{-\lambda n \hat{L}(h, S)\}]$ using standard techniques under the i.i.d. data-generation model: $S = (Z_1, \ldots, Z_n) \sim P_1^n$. Then with $Z \sim P_1$ and a few calculations (shown below in Appendix E.1) we obtain:

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

Assuming $M := \sup_h \mathbb{E}[\ell(h, Z)^2] < \infty$ (see Holland [2019] whose main result required this), using the function $f(h, s) = \lambda n \big(L(h) - \hat{L}(h, s)\big) - \frac{\lambda^2 n}{2} M$, with a fixed 'data-free' prior $Q^0$ the exponential moment $\xi = \mathbb{E}^0[e^{f(S,H)}]$ (i.e. $\xi = P_1^n \otimes Q^0[e^f]$) satisfies $\xi \leq 1$. This way we obtain the following PAC-Bayes type of bound under unbounded ('free range') losses:

**Theorem 15** *For any $n$, for any $P_1 \in \mathcal{M}_1(\mathcal{Z})$, for any data-free $Q^0 \in \mathcal{M}_1(\mathcal{H})$, for any loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, \infty)$, for any $Q \in \mathcal{K}(\mathcal{S}, \mathcal{H})$, for any $\lambda \in (0, \infty)$, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over size-$n$ i.i.d. samples $S \sim P_1^n$ we have*

$$Q_S[L] \leq Q_S[\hat{L}_S] + \frac{\mathrm{KL}(Q_S \| Q^0) + \log(1/\delta)}{n\lambda} + \frac{\lambda}{2} \sup_h \mathbb{E}[\ell(h, Z)^2] . \tag{22}$$

Essentially, this bound is of the form $Q_S[L] - Q_S[\hat{L}_S] \leq B/(n\lambda) + \lambda M/2$. With the optimal choice of $\lambda$ we get $Q_S[L] - Q_S[\hat{L}_S] \leq 2\sqrt{BM/(2n)}$, which gives a slow convergence rate of $O(1/\sqrt{n})$. The assumption of finite $M$ is satisfied e.g. when the loss is sub-gaussian or sub-exponential. It would be interesting to characterize all cases when $M < \infty$ holds. However, this simple bound illustrates that PAC-Bayes bounds are possible with unbounded loss functions.

### E.1 The calculations to bound the exponential moment

We start by calculating $\mathbb{E}[\exp\{-\lambda n \hat{L}(h, S)\}]$ with fixed $\lambda > 0$ and $h \in \mathcal{H}$. This means that the expectation is with respect to $S = (Z_1, \ldots, Z_n) \sim P_1^n$. By independence, and using the inequality $e^x \leq 1 + x + x^2/2$ valid for $x \leq 0$, we have

$$\begin{aligned}
\mathbb{E}[\exp\{-\lambda n \hat{L}(h, S)\}] &= \prod_{i \in [n]} \mathbb{E}[\exp\{-\lambda \ell(h, Z_i)\}] \\
&\leq \prod_{i \in [n]} \mathbb{E}[1 - \lambda \ell(h, Z_i) + \frac{\lambda^2}{2} \ell(h, Z_i)^2] \\
&= \prod_{i \in [n]} \Big(1 - \lambda \, \mathbb{E}[\ell(h, Z_i)] + \frac{\lambda^2}{2} \, \mathbb{E}[\ell(h, Z_i)^2]\Big)
\end{aligned}$$

and then using $1 + x \leq e^x$, which is valid for all $x$, the above is

$$\begin{aligned}
&\leq \prod_{i \in [n]} \exp\big\{-\lambda \, \mathbb{E}[\ell(h, Z_i)] + \frac{\lambda^2}{2} \, \mathbb{E}[\ell(h, Z_i)^2]\big\} \\
&= \exp\big\{-\lambda n L(h) + \frac{\lambda^2 n}{2} \, \mathbb{E}[\ell(h, Z)^2]\big\} .
\end{aligned}$$

In the last line we have used the identical distribution of the $Z_i$'s, namely $Z_i \sim P_1$, and a generic identical copy $Z \sim P_1$. Then, rearranging, we get as claimed that

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

These kinds of calculations are well known, however, we would like to acknowledge the section 'alternative proofs' of Thiemann [2016]. Then under the assumption $M = \sup_h \mathbb{E}[\ell(h, Z)^2] < \infty$, using the function $f(h, s) = \lambda n \big(L(h) - \hat{L}(h, s)\big) - \frac{\lambda^2 n}{2} M$ and a fixed 'data-free' prior $Q^0$, the exponential moment $\xi$ of $f$ under the joint distribution $P_1^n \otimes Q^0$ (i.e. $\xi = P_1^n[Q^0[e^f]]$) satisfies $\xi = \xi_{\mathrm{swap}} = Q^0[P_1^n[e^f]]$ (see discussion after Theorem 2 in Section 3), while the above calculations show that the latter satisfies $\xi_{\mathrm{swap}} \leq 1$.