PAC-Bayes under potentially heavy tails

Matthew J. Holland
Institute of Scientific and Industrial Research
Osaka University
matthew-h@ar.sanken.osaka-u.ac.jp

Abstract

We derive PAC-Bayesian learning guarantees for heavy-tailed losses, and obtain a novel optimal Gibbs posterior which enjoys finite-sample excess risk bounds at logarithmic confidence. Our core technique itself makes use of PAC-Bayesian inequalities in order to derive a robust risk estimator, which by design is easy to compute. In particular, only assuming that the first three moments of the loss distribution are bounded, the learning algorithm derived from this estimator achieves nearly sub-Gaussian statistical error, up to the quality of the prior.

1 Introduction

More than two decades ago, the origins of PAC-Bayesian learning theory were developed with the goal of strengthening traditional PAC learning guarantees[1] by explicitly accounting for prior knowledge [20, 15, 7]. Subsequent work developed finite-sample risk bounds for "Bayesian" learning algorithms which specify a distribution over the model [16]. These bounds are controlled using the empirical risk and the relative entropy between "prior" and "posterior" distributions, and hold uniformly over the choice of the latter, meaning that the guarantees hold for data-dependent posteriors, hence the naming. Furthermore, choosing the posterior to minimize PAC-Bayesian risk bounds leads to practical learning algorithms which have seen numerous successful applications [3].

Following this framework, a tremendous amount of work has been done to refine, extend, and apply the PAC-Bayesian framework to new learning problems. Tight risk bounds for bounded losses are due to Seeger [18] and Maurer [14], with the former work applying them to Gaussian processes. Bounds constructed using the loss variance in a Bernstein-type inequality were given by Seldin et al. [19], with a data-dependent extension derived by Tolstikhin and Seldin [21]. As stated by McAllester [17], virtually all the bounds derived in the original PAC-Bayesian theory "only apply to bounded loss functions." This technical barrier was solved by Alquier et al. [3], who introduce an additional error term depending on the concentration of the empirical risk about the true risk. This technique was subsequently applied to the log-likelihood loss in the context of Bayesian linear regression by Germain et al. [12], and further systematized by Bégin et al. [5]. While this approach lets us deal with unbounded losses, naturally the statistical error guarantees are only as good as the confidence intervals available for the empirical mean deviations. In particular, strong assumptions on all of the moments of the loss are essentially unavoidable using the traditional tools espoused by Bégin et al. [5], which means the "heavy-tailed" regime cannot be handled, where all we assume is that a few higher-order moments are finite (say finite variance and/or finite kurtosis). A new technique for deriving PAC-Bayesian bounds even under heavy-tailed losses is introduced by Alquier and Guedj [2]; their lucid procedure provides error rates even under heavy tails, but as the authors recognize, the guarantees are sub-optimal at high confidence levels due to direct dependence on the empirical risk, leading in turn to sub-optimal algorithms derived from these bounds.[2]

[1] PAC: Probably approximately correct [22].

[2] See work by Catoni [9], Devroye et al. [11] and the references within for background on the fundamental limitations of the empirical mean for real-valued random variables.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

In this work, while keeping many core ideas of Bégin et al. [5] intact, using a novel approach we obtain exponential tail bounds on the excess risk using PAC-Bayesian bounds that hold even under heavy-tailed losses. Our key technique is to replace the empirical risk with a new mean estimator inspired by the dimension-free estimators of Catoni and Giulini [10], designed to be computationally convenient. We review some key theory in section 2 before introducing the new estimator in section 3. In section 4 we apply this estimator to the PAC-Bayes setting, deriving a new robust optimal Gibbs posterior. Empirical inquiries into the properties of the new mean estimator are given in section 5. All proofs are relegated to supplementary materials.

2 PAC-Bayesian theory based on the empirical mean

Let us begin by briefly reviewing the best available PAC-Bayesian learning guarantees under general losses. Denote by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathcal{Z}$ a sequence of independent observations distributed according to common distribution μ. Denote by $\mathcal{H}$ a model/hypothesis class, from which the learner selects a candidate based on the n-sized sample. The quality of this choice can be measured in a pointwise fashion using a loss function $l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, assumed to be $l \geq 0$. The learning task is to achieve a small risk, defined by $R(h) \coloneqq \mathbf{E}_\mu \, l(h; \boldsymbol{z})$. Since the underlying distribution is inherently unknown, the canonical proxy is

$$\widehat{R}(h) \coloneqq \frac{1}{n} \sum_{i=1}^{n} l(h; \boldsymbol{z}_i), \quad h \in \mathcal{H}.$$

Let ν and ρ respectively denote "prior" and "posterior" distributions on the model $\mathcal{H}$. The so-called Gibbs risk induced by ρ, as well as its empirical counterpart are given by

$$G_\rho \coloneqq \mathbf{E}_\rho \, R = \int_{\mathcal{H}} R(h) \, d\rho(h), \quad \widehat{G}_\rho \coloneqq \mathbf{E}_\rho \, \widehat{R} = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{H}} l(h; \boldsymbol{z}_i) \, d\rho(h).$$

When our losses are almost surely bounded, lucid guarantees are available.

Theorem 1 (PAC-Bayes under bounded losses [16, 5])**.** *Assume $0 \leq l \leq 1$, and fix any arbitrary prior ν on $\mathcal{H}$. For any confidence level $\delta \in (0, 1)$, we have with probability no less than $1 - \delta$ over the draw of the sample that*

$$G_\rho \leq \widehat{G}_\rho + \sqrt{\frac{\boldsymbol{K}(\rho; \nu) + \log(2\sqrt{n}\delta^{-1})}{2n}}$$

uniformly in the choice of ρ.

Since the "good event" where the inequality in Theorem 1 holds is valid for any choice of ρ, the result holds even when ρ depends on the sample, which justifies calling it a posterior distribution. Optimizing this upper bound with respect to ρ leads to the so-called optimal Gibbs posterior, which takes a form which is readily characterized (cf. Remark 13).

The above results fall apart when the loss is unbounded, and meaningful extensions become challenging when exponential moment bounds are not available. As highlighted in section 1 above, over the years, the analytical machinery has evolved to provide general-purpose PAC-Bayesian bounds even under heavy-tailed data. The following theorem of Alquier and Guedj [2] extends the strategy of Bégin et al. [5] to obtain bounds under the weakest conditions we know of.

Theorem 2 (PAC-Bayes under heavy-tailed losses [2])**.** *Take any $p > 1$ and set $q = p/(p-1)$. For any confidence level $\delta \in (0, 1)$, we have with probability no less than $1 - \delta$ over the draw of the sample that*

$$G_\rho \leq \widehat{G}_\rho + \left(\frac{\mathbf{E}_\nu \, |\widehat{R} - R|^q}{\delta} \right)^{\frac{1}{q}} \left(\int_{\mathcal{H}} \left(\frac{d\rho}{d\nu} \right)^p d\nu \right)^{\frac{1}{p}}$$

uniformly in the choice of ρ.

For concreteness, consider the case of $p = 2$, where $q = 2/(2-1) = 2$, and assume that the variance of the loss is $\operatorname{var}_\mu l(h; \boldsymbol{z})$ is ν-finite, namely that

$$V_\nu \coloneqq \int_{\mathcal{H}} \operatorname{var}_\mu l(h; \boldsymbol{z}) \, d\nu(h) < \infty.$$

From Proposition 4 of Alquier and Guedj [2], we have $\mathbf{E}_\nu |\widehat{R} - R|^2 \leq V_\nu / n$. It follows that on the high-probability event, we have

$$G_\rho \leq \widehat{G}_\rho + \sqrt{\frac{V_\nu}{n\,\delta} \left(\int_{\mathcal{H}} \left(\frac{d\rho}{d\nu} \right)^2 d\nu \right)}$$

While the $\sqrt{n}$ rate and dependence on a divergence between ν and ρ are similar, note that the dependence on the confidence level $\delta \in (0,1)$ is polynomial; compare this with the logarithmic dependence available in Theorem 1 above when the losses were bounded.

For comparison, our main result of section 4 is a uniform bound on the Gibbs risk: with probability no less than $1 - \delta$, we have

$$G_\rho \leq \widehat{G}_{\rho,\psi} + \frac{1}{\sqrt{n}} \left(\boldsymbol{K}(\rho;\nu) + \frac{\log(8\pi M_2 \delta^{-2})}{2} + M_2 + \nu_n^*(\mathcal{H}) - 1 \right) + O\left(\frac{1}{n} \right)$$

where $\widehat{G}_{\rho,\psi}$ is an estimator of G_ρ defined in section 3, $\nu_n^*(\mathcal{H})$ is a term depending on the quality of prior ν, and the key constants are bounds such that for all $h \in \mathcal{H}$ we have $M_2 \geq \mathbf{E}_\mu\, l(h;\boldsymbol{z})^2$. As long as the first three moments are finite, this guarantee holds, and thus both sub-Gaussian and heavy-tailed losses (e.g., with infinite higher-order moments) are permitted. Given any valid M_2, the PAC-Bayesian upper bound above can be minimized in ρ based on the data, and thus an optimal Gibbs posterior can also be computed in practice. In section 4, we characterize this "robust posterior."

3 A new estimator using smoothed Bernoulli noise

Notation In this section, we are dealing with the specific problem of robust mean estimation, thus we specialize our notation here slightly. Data observations will be $x_1, \ldots, x_n \in \mathbb{R}$, assumed to be independent copies of $x \sim \mu$. Denote the index set $[k] \coloneqq \{1, 2, \ldots, k\}$. Write $\mathcal{M}_+^1(\Omega, \mathcal{A})$ for the set of all probability measures defined on the measurable space $(\Omega, \mathcal{A})$. Write $\boldsymbol{K}(P, Q)$ for the relative entropy between measures P and Q (also known as the KL divergence; definition in appendix). We shall typically suppress $\mathcal{A}$ and even Ω in the notation when it is clear from the context. Let ψ be a bounded, non-decreasing function such that for some $b > 0$ and all $u \in \mathbb{R}$,

$$-\log\left(1 - u + u^2/b\right) \leq \psi(u) \leq \log\left(1 + u + u^2/b\right). \tag{1}$$

As a concrete and analytically useful example, we shall use the piecewise polynomial function of Catoni and Giulini [10], defined by

$$\psi(u) \coloneqq \begin{cases} u - u^3/6, & -\sqrt{2} \leq u \leq \sqrt{2} \\ 2\sqrt{2}/3, & u > \sqrt{2} \\ -2\sqrt{2}/3, & u < -\sqrt{2} \end{cases} \tag{2}$$

which for $b = 2$ satisfies (1). Slightly looser bounds hold with $b = 1$ for an analogous procedure using a Huber-type influence function.

Estimator definition We consider a straightforward procedure, in which the data are subject to a soft truncation after re-scaling, defined by

$$\widehat{x} \coloneqq \frac{s}{n} \sum_{i=1}^n \psi\left(\frac{x_i}{s} \right) \tag{3}$$

where $s > 0$ is a re-scaling parameter. Depending on the setting of s, this function can very closely approximate the sample mean, and indeed modifying this scaling parameter controls the bias of this estimator in a direct way, which can be quantified as follows. As the scale grows, note that

$$s\psi\left(\frac{x}{s}\right) = x - \frac{x^3}{6s^2} \to x, \quad \text{as } s \to \infty$$

which implies that taking expectation with respect to the sample and $s \to \infty$, in the limit this estimator is unbiased, with

$$\mathbf{E}\left(\frac{s}{n} \sum_{i=1}^n \psi\left(\frac{x_i}{s} \right) \right) = \mathbf{E}_\mu\, x - \frac{\mathbf{E}_\mu\, x^3}{6s^2} \to \mathbf{E}_\mu\, x.$$

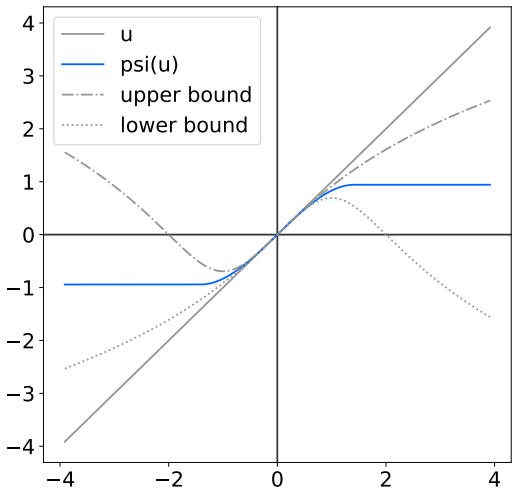


Figure 1: Graph of the Catoni function $\psi(u)$ over $\pm\sqrt{2} \pm 2.5$.

On the other hand, taking s closer to zero implies that more observations will be truncated. Taking s small enough,[3] we have

$$\frac{s}{n} \sum_{i=1}^{n} \psi\left(\frac{x_i}{s}\right) = \frac{2\sqrt{2}s}{3n} \left(|\mathcal{I}_+| - |\mathcal{I}_-|\right),$$

which converges to zero as $s \to 0$. Here the positive/negative indices are $\mathcal{I}_+ := \{i \in [n] : x_i > 0\}$ and $\mathcal{I}_- := \{i \in [n] : x_i < 0\}$. Thus taking s too small means that only the signs of the observations matter, and the absolute value of the estimator tends to become too small.

High-probability deviation bounds for $\widehat{x}$ We are interested in high-probability bounds on the deviations $|\widehat{x} - \mathbf{E}_\mu x|$ under the weakest possible assumptions on the underlying data distribution. To obtain such guarantees in a straightforward manner, we make the simple observation that the estimator $\widehat{x}$ defined in (3) can be related to an estimator with smoothed noise as follows. Let $\epsilon_1, \ldots, \epsilon_n$ be an iid sample of noise $\epsilon \in \{0, 1\}$ with distribution Bernoulli(θ) for some $0 < \theta < 1$. Then, taking expectation with respect to the noise sample, one has that

$$\widehat{x} = \frac{1}{\theta}\, \mathbf{E} \left(\frac{s}{n} \sum_{i=1}^{n} \psi\left(\frac{x_i\, \epsilon_i}{s}\right) \right). \tag{4}$$

This simple observation becomes useful to us in the context of the following technical fact.

Lemma 3. *Assume we are given some independent data $x_1, \ldots, x_n$, assumed to be copies of the random variable $x \sim \mu$. In addition, let $\epsilon_1, \ldots, \epsilon_n$ similarly be independent observations of "strategic noise," with distribution $\epsilon \sim \rho$ that we can design. Fix an arbitrary prior distribution ν, and consider $f : \mathbb{R}^2 \to \mathbb{R}$, assumed to be bounded and measurable. Write $\boldsymbol{K}(\rho; \nu)$ for the Kullback-Leibler divergence between distributions ρ and ν. It follows that with probability no less than $1 - \delta$ over the random draw of the sample, we have*

$$\mathbf{E} \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i, \epsilon_i) \right) \leq \int \log \mathbf{E}_\mu \exp(f(x, \epsilon))\, d\rho(\epsilon) + \frac{\boldsymbol{K}(\rho; \nu) + \log(\delta^{-1})}{n},$$

uniform in the choice of ρ, where expectation on the left-hand side is over the noise sample.

The special case of interest here is $f(x, \epsilon) = \psi(x\epsilon/s)$. Using (1) and Lemma 3, with prior $\nu =$ Bernoulli$(1/2)$ and posterior $\rho =$ Bernoulli(θ), it follows that on the $1 - \delta$ high-probability event,

[3] More precisely, taking $s \leq \min\{|x_i| : i \in [n]\}/\sqrt{2}$.

uniform in the choice of $0 < \theta < 1$, we have

$$\left(\frac{\theta}{s}\right)\widehat{x} \leq \int \left(\frac{\epsilon\,\mathbf{E}_\mu\, x}{s} + \frac{\epsilon^2\,\mathbf{E}_\mu\, x^2}{2s^2}\right) d\rho(\epsilon) + \frac{\boldsymbol{K}(\rho;\nu) + \log(\delta^{-1})}{n} \tag{5}$$
$$= \frac{\theta\,\mathbf{E}_\mu\, x}{s} + \frac{\theta\,\mathbf{E}_\mu\, x^2}{2s^2} + \frac{1}{n}\left(\theta\log(2\theta) + (1-\theta)\log(2(1-\theta)) + \log(\delta^{-1})\right)$$

where we have used the fact that $\mathbf{E}\,\epsilon^2 = \mathbf{E}\,\epsilon = \theta$ in the Bernoulli case. Dividing both sides by (θ/s) and optimizing this as a function of $s > 0$ yields a closed-form expression for s depending on the second moment, the confidence δ, and θ. Analogous arguments yield lower bounds on the same quantity. Taking these facts together, we have the following proposition, which says that assuming only finite second moments $\mathbf{E}_\mu\, x^2 < \infty$, the proposed estimator achieves exponential tail bounds scaling with the second non-central moment.

Proposition 4 (Concentration of deviations)**.** *Scaling with $s^2 = n\,\mathbf{E}_\mu\, x^2/2\log(\delta^{-1})$, the estimator defined in (3) satisfies*

$$|\widehat{x} - \mathbf{E}_\mu\, x| \leq \sqrt{\frac{2\,\mathbf{E}_\mu\, x^2 \log(\delta^{-1})}{n}} \tag{6}$$

with probability at least $1 - 2\delta$.

Remark 5. While the above bound (6) depends on the true second moment, the result is easily extended to hold for any valid upper bound on the moment, which is what will inevitably have to be used in practice.

Centered estimates Note that the bound (6) depends on the second moment of the underlying data; this is in contrast to M-estimators which due to a natural "centering" of the data typically have tail bounds depending on the variance [9]. This results in a sensitivity to the absolute value of the location of the distribution, e.g., on a distribution with unit variance and $\mathbf{E}_\mu\, x = 0$ will tend to be much better than a distribution with $\mathbf{E}_\mu\, x = 10^4$. Fortunately, a simple centering strategy works well to alleviate this sensitivity, as follows. Without loss of generality, assume that the first $0 < k < n$ estimates are used for constructing a shifting device, with the remaining $n - k > 0$ points left for running the usual routine on shifted data. More concretely, define

$$\bar{x}_\psi = \frac{\bar{s}}{k}\sum_{i=1}^{k} \psi\left(\frac{x_i}{\bar{s}}\right), \text{ where } \bar{s}^2 = \frac{k\,\mathbf{E}_\mu\, x^2}{2\log(\delta^{-1})}. \tag{7}$$

From (6) in Proposition 4, we have

$$|\bar{x}_\psi - \mathbf{E}_\mu\, x| \leq \varepsilon_k \coloneqq \sqrt{\frac{2\,\mathbf{E}_\mu\, x^2 \log(\delta^{-1})}{k}}$$

on an event with probability no less than $1 - 2\delta$, over the draw of the k-sized sub-sample. Using this, we shift the remaining data points as $x_i' \coloneqq x_i - \bar{x}_\psi$. Note that the second moment of this data is bounded as $\mathbf{E}_\mu(x')^2 \leq \mathrm{var}_\mu\, x + \varepsilon_k^2$. Passing these shifted points through (3) with analogous second moment bounds used for scaling, we have

$$\widehat{x}' = \frac{s}{(n-k)}\sum_{i=k+1}^{n} \psi\left(\frac{x_i'}{s}\right), \text{ where } s^2 = \frac{(n-k)(\mathrm{var}_\mu\, x + \varepsilon_k^2)}{2\log(\delta^{-1})}. \tag{8}$$

Shifting the resulting output back to the original location by adding and shifting $\widehat{x}'$ back to the original location by adding $\bar{x}_\psi$, conditioned on $\bar{x}_\psi$, we have by (6) again that

$$|(\widehat{x}' + \bar{x}_\psi) - \mathbf{E}_\mu\, x| = |\widehat{x} - \mathbf{E}_\mu(x - \bar{x}_\psi)| \leq \sqrt{\frac{2(\mathrm{var}_\mu\, x + \varepsilon_k^2)\log(\delta^{-1})}{n-k}}$$

with probability no less than $1 - 2\delta$ over the draw of the remaining $n - k$ points. Defining the centered estimator as $\widehat{x} = \widehat{x}' + \bar{x}_\psi$, and taking a union bound over the two "good events" on the independent sample subsets, we may thus conclude that

$$\mathbf{P}\left\{|\widehat{x} - \mathbf{E}_\mu\, x| > \varepsilon\right\} \leq 4\exp\left(\frac{-(n-k)\varepsilon^2}{2(\mathrm{var}_\mu\, x + \varepsilon_k^2)}\right) \tag{9}$$

where probability is over the draw of the full n-sized sample. While one takes a hit in terms of the sample size, the variance works to combat sensitivity to the distribution location (see section 5 for empirical tests).

4 PAC-Bayesian bounds for heavy-tailed data

An import and influential paper due to D. McAllester gave the following theorem as a motivating result. To get started, we give a slightly modified version of his result.

Theorem 6 (McAllester [15], Preliminary Theorem 2)**.** *Let ν be a prior probability distribution over $\mathcal{H}$, assumed countable, and to be such that $\nu(h) > 0$ for all $h \in \mathcal{H}$. Consider the pattern recognition task with $\boldsymbol{z} = (\boldsymbol{x}, y) \in \mathcal{X} \times \{-1, 1\}$, and the classification error $l(h; \boldsymbol{z}) = I\{h(\boldsymbol{x}) \neq y\}$. Then with probability no less than $1 - \delta$, for any choice of $h \in \mathcal{H}$, we have*

$$R(h) \leq \frac{1}{n} \sum_{i=1}^{n} l(h; \boldsymbol{z}_i) + \sqrt{\frac{\log\left(1/\nu(h)\right) + \log\left(1/\delta\right)}{2n}}$$

One quick glance at the proof of this theorem shows that the bounded nature of the observations plays a crucial role in deriving excess risk bounds of the above form, as it is used to obtain concentration inequalities for the empirical risk about the true risk. While analogous concentration inequalities hold under slightly weaker assumptions, when considering the potentially heavy-tailed setting, one simply cannot guarantee that empirical risk is tightly concentrated about the true risk, which prevents direct extensions of such theorems. With this in mind, we take a different approach, that does not require the empirical mean to be well-concentrated.

Our motivating pre-theorem The basic idea of our approach is very simple: instead of using the sample mean, bound the off-sample risk using a more robust estimator which is easy to compute directly, and which allows risk bounds even under unbounded, potentially heavy-tailed losses. Define a new approximation of the risk by

$$\widehat{R}_{\psi}(h) \coloneqq \frac{s}{n} \sum_{i=1}^{n} \psi\left(\frac{l(h; \boldsymbol{z}_i)}{s}\right), \tag{10}$$

for $s > 0$. Note that this is just a direct application of the robust estimator defined in (3) to the case of a loss which depends on the choice of candidate $h \in \mathcal{H}$. As a motivating result, we basically re-prove McAllester's result (Theorem 6) under much weaker assumptions on the loss, using the statistical properties of the new risk estimator (10), rather than relying on classical Chernoff inequalities.

Theorem 7 (Pre-theorem)**.** *Let ν be a prior probability distribution over $\mathcal{H}$, assumed countable. Assume that $\nu(h) > 0$ for all $h \in \mathcal{H}$, and that $m_2(h) \coloneqq \mathbf{E}\, l(h; \boldsymbol{z})^2 < \infty$ for all $h \in \mathcal{H}$. Setting the scale in (10) to $s_h^2 = n\, m_2(h)/2\log(\delta^{-1})$, then with probability no less than $1 - 2\delta$, for any choice of $h \in \mathcal{H}$, we have*

$$R(h) \leq \widehat{R}_{\psi}(h) + \sqrt{\frac{2m_2(h)\left(\log(1/\nu(h)) + \log(1/\delta)\right)}{n}}.$$

Remark 8. We note that all quantities on the right-hand side of Theorem 7 are easily computed based on the sample, except for the second moment m_2, which in practice must be replaced with an empirical estimate. With an empirical estimate of m_2 in place, the upper bound can easily be used to derive a learning algorithm.

Uncountable model case Next we extend the previous motivating theorem to a more general result on a potentially uncountable $\mathcal{H}$, using stochastic learning algorithms, as has become standard in the PAC-Bayes literature. We need a few technical conditions, listed below:

1. Bounds on lower-order moments. For all $h \in \mathcal{H}$, we require $\mathbf{E}_{\mu}\, l(h; \boldsymbol{z})^2 \leq M_2 < \infty$, $\mathbf{E}_{\mu}\, l(h; \boldsymbol{z})^3 \leq M_3 < \infty$.
2. Bounds on the risk. For all $h \in \mathcal{H}$, we require $R(h) \leq \sqrt{nM_2/(4\log(\delta^{-1}))}$.
3. Large enough confidence. We require $\delta \leq \exp(-1/9) \approx 0.89$.

These conditions are quite reasonable, and easily realized under heavy-tailed data, with just lower-order moment assumptions on μ and say a compact class $\mathcal{H}$. The new terms that appear in our bounds that do no appear in previous works are $\widehat{G}_{\rho,\psi} \coloneqq \mathbf{E}_{\rho}\, \widehat{R}_{\psi}$ and $\nu_n^*(\mathcal{H}) = \mathbf{E}_{\nu} \exp(\sqrt{n}(R - \widehat{R}_{\psi}))/\, \mathbf{E}_{\nu} \exp(R - \widehat{R}_{\psi})$. The former is the expectation of the proposed robust estimator with respect to posterior ρ, and the latter is a term that depends directly on the quality of the prior ν.

Theorem 9. *Let ν be a prior distribution on model $\mathcal{H}$. Let the three assumptions listed above hold. Setting the scale in (10) to $s^2 = n\,M_2/2\log(\delta^{-1})$, then with probability no greater than $1-\delta$ over the random draw of the sample, it holds that*

$$G_\rho \leq \widehat{G}_{\rho,\psi} + \frac{1}{\sqrt{n}}\left(\boldsymbol{K}(\rho;\nu) + \frac{\log(8\pi M_2\delta^{-2})}{2} + M_2 + \nu_n^*(\mathcal{H}) - 1\right) + O\left(\frac{1}{n}\right)$$

for any choice of probability distribution ρ on $\mathcal{H}$, since $G_\rho < \infty$ by assumption.

Remark 10. As is evident from the statement of Theorem 9, the convergence rate is clear for all terms but $\nu_n^*(\mathcal{H})/\sqrt{n}$. In our proof, we use a modified version of the elegant and now-standard strategy formulated by Bégin et al. [5]. A glance at the proof shows that under this strategy, there is essentially no way to avoid dependence on $\nu_n^*(\mathcal{H})$. Since the random variable $R - \widehat{R}_\psi$ is bounded over the random draw of the sample and $h \sim \nu$, the bounds still hold and are non-trivial. That said, $\nu_n^*(\mathcal{H})$ may indeed increase as $n \to \infty$, potentially spoiling the $\sqrt{n}$ rate, and even consistency in the worst case. Clearly $\nu_n^*(\mathcal{H})$ presents no troubles if $R - \widehat{R}_\psi \leq 0$ on a high-probability event, but note that this essentially amounts to asking for a prior that on average realizes bounds that are better than we can guarantee for *any* posterior though the above analysis. Such a prior may indeed exist, but if it were known, then that would eliminate the need for doing any learning at all. If the deviations $R - \widehat{R}_\psi$ are truly sub-Gaussian [6], then the $\sqrt{n}$ rate can be easily obtained. However, impossibility results from Devroye et al. [11] suggest that under just a few finite moment assumptions, such an estimator cannot be constructed. As such, here we see a clear limitation of the established PAC-Bayes analytical framework under potentially heavy-tailed data. Since the change of measures step in the proof is fundamental to the basic argument, it appears that concessions will have to be made, either in the form of slower rates, deviations larger than the relative entropy, or weaker dependence on $1/\delta$.

Remark 11. Note that while in its tightest form, the above bound requires knowledge of $\mathbf{E}_\mu\, l(h;\boldsymbol{z})^2$, we may set $s > 0$ used to define $\widehat{R}_\psi$ using any valid upper bound M_2, under which the above bound still holds as-is, using known quantities. Furthermore, for reference the content of the $O(1/n)$ term in the above bound takes the form

$$\frac{1}{n}\left(2\sqrt{V\log(\delta^{-1})} + \frac{M_3\log(\delta^{-1})}{3M_2\sqrt{n}}\right)$$

where V is an upper bound on the variance $\operatorname{var}_\mu l(h;\boldsymbol{z}) \leq V < \infty$ over $h \in \mathcal{H}$.

As a principled approach to deriving stochastic learning algorithms, one naturally considers the choice of posterior ρ in Theorem 9 that minimizes the upper bound. This is typically referred to as the optimal Gibbs posterior [12], and takes a form which is easily characterized, as we prove in the following proposition.

Proposition 12 (Robust optimal Gibbs posterior)**.** *The upper bound of Theorem 9 is optimized by a data-dependent posterior distribution $\widehat{\rho}$, defined in terms of its density function with respect to the prior ν as*

$$\left(\frac{d\widehat{\rho}}{d\nu}\right)(h) = \frac{\exp\left(-\sqrt{n}\widehat{R}_\psi(h)\right)}{\mathbf{E}_\nu \exp\left(-\sqrt{n}\widehat{R}_\psi\right)}.$$

Furthermore, the risk bound under the optimal Gibbs posterior takes the form

$$G_{\widehat{\rho}} \leq \frac{1}{\sqrt{n}}\left(\log \mathbf{E}_\nu \exp\left(\sqrt{n}\widehat{R}_\psi\right) + \frac{\log(8\pi M_2\delta^{-1})}{2} + M_2 + \nu_n^*(\mathcal{H}) - 1\right) + O\left(\frac{1}{n}\right)$$

with probability no less than $1-\delta$ over the draw of the sample.

Remark 13 (Comparison with traditional Gibbs posterior). In traditional PAC-Bayes analysis [12, Equation 8], the optimal Gibbs posterior, let us write $\widehat{\rho}_{\text{emp}}$, is defined by

$$\left(\frac{d\widehat{\rho}_{\text{emp}}}{d\nu}\right)(h) = \frac{\exp\left(-n\widehat{R}(h)\right)}{\mathbf{E}_\nu \exp\left(-n\widehat{R}\right)}$$

where $\widehat{R}(h) = n^{-1}\sum_{i=1}^n l(h;\boldsymbol{z}_i)$ is the empirical risk. We have $n\widehat{R}$ and $\sqrt{n}\widehat{R}_\psi$, but since scaling in the latter case should be done with $s \propto \sqrt{n}$, so in both cases the $1/n$ factor cancels out. In the special

case of the negative log-likelihood loss, Germain et al. [12] demonstrate that the optimal Gibbs posterior coincides with the classical Bayesian posterior. As noted by Alquier et al. [3], the optimal Gibbs posterior has shown strong empirical performance in practice, and variational approaches have been proposed as efficient alternatives to more traditional MCMC-based implementations. Comparison of both the computational and learning efficiency of our proposed "robust Gibbs posterior" with the traditional Gibbs posterior is a point of significant interest moving forward.

5 Empirical analysis

In this section, we use tightly controlled simulations to investigate how the performance of $\widehat{x}$ (cf. (3) and Proposition 4) compares with the sample mean and other robust estimators. We pay particular attention to how performance depends on the underlying distribution family, the value of second moments, and the sample size.

Experimental setup For each experimental setting and each independent trial, we generate a sample $x_1, \ldots, x_n$ of size n, compute some estimator $\{x_i\}_{i=1}^n \mapsto \widehat{x}$, and record the deviation $|\widehat{x} - \mathbf{E}_\mu|$. The sample sizes range over $n \in \{10, 20, 30, \ldots, 100\}$, and the number of trials is 10^4. We draw data from two distribution families, the Normal family with mean a and variance b^2, and the log-Normal family, with log-mean $a_{\log}$ and log-variance $b_{\log}^2$, under multiple parameter settings. In particular, we consider the impact of shifting the distribution location over $[-40.0, 40.0]$, with small and large variance settings. Regarding the variance, we have "low," "mid," and "high" settings, which correspond to $b = 0.5, 5.0, 50.0$ in the Normal case, and $b_{\log} = 1.1, 1.35, 1.75$ in the log-Normal case. Over all settings, the log-location parameter of the log-Normal data is fixed at $a_{\log} = 0$. Shifting the Normal data is trivially accompished by taking the desired $a \in [-40.0, 40.0]$. Shifting the log-Normal data is accomplished by subtracting the true mean (pre-shift) equal to $\exp(a_{\log} + b_{\log}^2/2)$ to center the data, and subsequently adding the desired location.

The methods being compared are as follows: `mean` denotes the empirical mean, `med` the empirical median,[4] `mult_g` is the estimator of Holland [13] using smoothed Gaussian noise, `mult_b` the proposed estimator $\widehat{x}$ defined in (3) using smoothed Bernoulli noise, and finally `mult_bc` the *centered* version of $\widehat{x}$, see the discussion culminating in (9). The latter methods are given access to the true variance or second moment as needed for scaling purposes, and all algorithms are run with confidence parameter $\delta = 0.01$.

Impact of distribution family In Figure 2, we give histograms of the deviations for each method of interest under high variance settings. Colored vertical rules correspond to the error bounds for $\widehat{x}$ under Gaussian noise and Bernoulli noise (bound via Proposition 4), with probability δ. When the standard deviation is not much larger than the mean, we can see substantial improvement over traditional estimators. The bias introduced by the different $\widehat{x}$ choices is clearly far smaller on average than the median, with substantially improved sensitivity to outliers when compared with the mean. The centered version of $\widehat{x}$ has a deviation distribution somewhere between that of the empirical mean and that of the other $\widehat{x}$ choices.

Impact of distribution location In Figure 3 (a), we plot the graph of average/median deviations over trials, taken as a function of the true location $\mathbf{E}_\mu x$. From these results, two clear observations can be made. First, note that the performance of the Gaussian-type (`mult_g`) and Bernoulli-type (`mult_b`) estimators methods tend to differ greatly as a function of the true mean; in particular, we see that the bias of the Gaussian case is far more sensitive to the true location, providing strong evidence for use of our proposed Bernoulli version, which is less expensive, essentially uniformly better than the Gaussian version (as we would expect from the tighter bounds), with error growing slower as a function of the true mean value. Second, the fact that the centering procedure works very well to mitigate the effect of the second moment value is lucid, also a price is paid in overall accuracy due to the naive sample-splitting technique discussed used.

Impact of sample size In Figure 3 (b), we show the graph of average/median deviations taken over all trials, viewed as a function of the sample size n. The most distinct observation that can be made

[4] After sorting, this is computed as the middle point when n is odd, or the average of the two middle points when n is even.

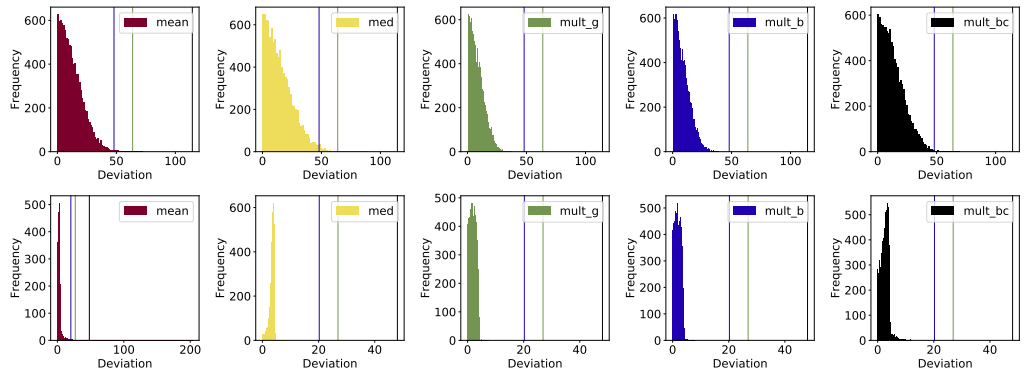
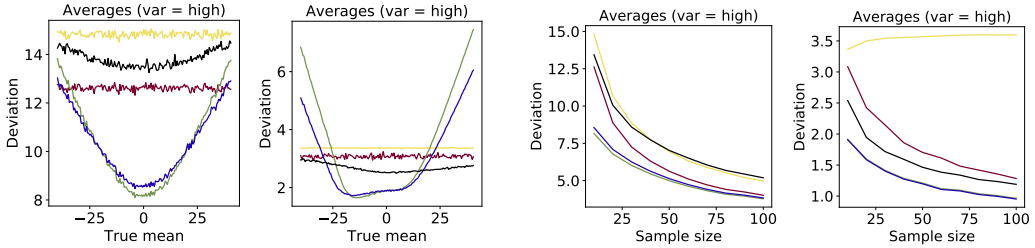


Figure 2: Histograms of deviations $|\widehat{x} - \mathbf{E}_\mu x|$ for different distributions and estimators, with accompanying error bounds. Sample size is $n = 10$. Distributions centered such that mean is equal to "low" level standard deviation. Top: Normal data. Bottom: log-Normal data.



(a) (b)

Figure 3: (a) Deviations $|\widehat{x} - \mathbf{E}_\mu x|$ as a function of the true mean $\mathbf{E}_\mu x$. (b) Deviations $|\widehat{x} - \mathbf{E}_\mu x|$ as a function of the sample size n. In both sub-figures, left is Normal data, right is log-Normal data.

here is that the estimator $\widehat{x}$ (3) considered here has learning efficiency which is far superior to the empirical mean and median, though as expected the centered version of $\widehat{x}$ has poorer efficiency, a direct result of the sample-splitting scheme used in its definition. As discussed before, this comes with the caveat that the mean cannot be too much larger than the standard deviation; when the second moment is exceedingly large, this leads to a rather large bias as seen in Figure 3 (a) previously.

6 Conclusions

The main contribution of this paper was to develop a novel approach to obtaining PAC-Bayesian learning guarantees, which admits deviations with exponential tails under weak moment assumptions on the underlying loss distribution, while still being computationally amenable. In this work, our chief interest was the fundamental problem of obtaining strong guarantees for stochastic learning algorithms which can reflect prior knowledge about the data-generating process, from which we derived a new robust Gibbs posterior. Moving forward, a deeper study of the statistical nature of this new stochastic learning algorithm, as well as computational considerations to be made in practice are of significant interest.

Acknowledgments

This work was partially supported by the JSPS KAKENHI Grant Number 18H06477.

A Technical appendix

A.1 Preparatory results

Relative entropy Here we recall the basic notions of the relative entropy, or Kullback-Leibler divergence, between two probability distributions. Consider P and Q, both defined over a finite space Ω. The relative entropy of P from Q is defined

$$\boldsymbol{K}(P;Q) \coloneqq \sum_{\omega\in\Omega} P(\omega)\log\left(\frac{P(\omega)}{Q(\omega)}\right), \tag{11}$$

where this definition clearly includes the possibility that $\boldsymbol{K}(P;Q)=\infty$, which occurs only when Q assigns zero probability to an element that P assigns positive probability to.

More generally, when Ω is potentially uncountably infinite, consider two probabilities P and Q on the measurable space $(\Omega,\mathcal{A})$, where $\mathcal{A}$ is an appropriate σ-algebra.[5] In this case, the relative entropy is defined

$$\boldsymbol{K}(P;Q) \coloneqq \int_\Omega \log\left(\frac{dP}{dQ}\right) dP, \qquad P \ll Q \tag{12}$$

where dP/dQ denotes the Radon-Nikodym derivative of P with respect to Q, typically called the density of P with respect to Q. The basic underlying technical assumption, denoted $P \ll Q$, is that P be absolutely continuous with respect to Q, meaning that $P(A)=0$ whenever $Q(A)=0$, for $A\in\mathcal{A}$. In the event that $P\ll Q$ does not hold, by convention we define $\boldsymbol{K}(P;Q)\coloneqq\infty$. Recall that the Radon-Nikodym theorem guarantees that when $P\ll Q$, there exists a measurable function $g\geq 0$ such that

$$P(A)=\int_A g\,dQ, \qquad A\in\mathcal{A}.$$

This function g is unique in the sense that if there exists another f satisfying the above equality, then $f=g$ almost everywhere $[Q]$. This uniqueness justifies using the notation dP/dQ, and calling this function *the* density of P (rather than *a* density of P).

Lemma 14 (Chain rule)**.** *On measure space $(\Omega,\mathcal{A},Q)$, let $g\geq 0$ be a Borel-measurable function, and define measure P by*

$$P(A)=\int_A g\,dQ, \qquad A\in\mathcal{A}.$$

For any Borel-measurable function f on Ω, it follows that

$$\int_\Omega f\,dP=\int_\Omega fg\,dQ.$$

Proof. See section 2.2, problem 4 of Ash and Doleans-Dade [4]. □

Lemma 15 (Non-negativity of relative entropy)**.** *For any probabilities P and Q, we have $\boldsymbol{K}(P;Q)\geq 0$.*

Proof of Lemma 15. If $P\ll Q$ does not hold, then $\boldsymbol{K}(P;Q)=\infty$ and non-negativity follows trivially. As for the case of $P\ll Q$, we begin with the basic logarithmic inequality $x<(1+x)\log(1+x)$ for any $x>-1$ [1]. We thus have $x-1<x\log(x)$ for any $x>0$. Using this inequality and the chain rule (Lemma 14), we have

$$\begin{aligned}
\boldsymbol{K}(P;Q) &= \mathbf{E}_P \log\frac{dP}{dQ} \\
&= \mathbf{E}_Q \frac{dP}{dQ}\log\frac{dP}{dQ} \\
&\geq \mathbf{E}_Q\left(\frac{dP}{dQ}-1\right) \\
&= 0.
\end{aligned}$$

The final equality uses the Radon-Nikodym theorem. □

[5] A certain degree of measure theory is assumed in this exposition, at approximately the level of the first few chapters of Ash and Doleans-Dade [4].

Lemma 16 (Lower bound on Bernoulli relative entropy)**.** *The relative entropy between Bernoulli(p) and Bernoulli(q) is bounded below by $\boldsymbol{K}(p;q) \geq 2(p-q)^2$.*

Proof of Lemma 16. Consider the function $f(p,q)$ defined

$$f(p,q) \coloneqq \boldsymbol{K}(p;q) - 2(p-q)^2.$$

Fix any arbitrary $p \in (0,1)$, and take the derivative with respect to q, noting that

$$\frac{d}{dq} f(p,q) = (-1)(p-q)\left(\frac{1}{q(1-q)} - 4\right).$$

Using the basic fact that $q(1-q) \leq 1/4$ for all $q \in (0,1)$, we have that the factor $(q(1-q))^{-1} - 4$ is non-negative. Thus, the slope is negative when $p > q$, postive when $p < q$, and zero when $p = q$. Thus this is the only minimum of the function in q. Note that $f(p,p) = 0$, and so for all $q \in (0,1)$ it follows that $f(p,q) \geq 0$. This holds for any choice of p as well, implying the desired result by the definition of f. □

Lemma 17 (Chernoff bound for Bernoulli data)**.** *Let $x_1, \ldots, x_n$ be independent and identically distributed random variables, taking values $x \in \{0,1\}$. Write $\bar{x} \coloneqq n^{-1}\sum_{i=1}^n x_i$ for the sample mean. The tails of the sample mean deviations can be bounded as*

$$\mathbf{P}\left\{\bar{x} - \mathbf{E}\, x > \varepsilon\right\} \leq \exp\left(-2n\varepsilon^2\right)$$
$$\mathbf{P}\left\{\bar{x} - \mathbf{E}\, x < -\varepsilon\right\} \leq \exp\left(-2n\varepsilon^2\right)$$

for any $0 < \varepsilon < 1 - \mathbf{E}\, x$.

Proof of Lemma 17. For random variable $x \sim$ Bernoulli(θ), recall that using Markov's inequality, for any $t > 0$ we have

$$\begin{aligned}\mathbf{P}\{X > \varepsilon\} &= \mathbf{P}\{\exp(tX) > \exp(t\varepsilon)\} \\ &\leq \exp(-t\varepsilon)\, \mathbf{E}\, e^{tX} \\ &= \exp(-t\varepsilon)\left(1 - \theta + \theta e^t\right).\end{aligned}$$

Taking the derivative of this upper bound with respect to t and setting it to zero, we obtain the condition

$$t^*(\varepsilon) = \log\left(\frac{\varepsilon}{\theta}\right)\left(\frac{1-\theta}{1-\varepsilon}\right),$$

where we write $t^*(\varepsilon)$ to emphasize the dependence of t^* on ε. We must have $t^*(\varepsilon) > 0$ for the bounds to hold. The value being passed into the $\log$ function must be greater than one. Fortunately, some simple re-arranging of factors shows that

$$\left(\frac{\varepsilon}{\theta}\right)\left(\frac{1-\theta}{1-\varepsilon}\right) > 1 \iff \varepsilon > \theta.$$

So we have $t^*(\varepsilon) > 0$ whenever $\theta < \varepsilon < 1$. Plugging this in, some algebra shows that

$$\begin{aligned}\exp(-t^*\varepsilon)\left(1 - \theta + \theta e^{t^*}\right) &= \exp\left((1-\varepsilon)\log\left(\frac{1-\theta}{1-\varepsilon}\right) + \varepsilon\log\left(\frac{\theta}{\varepsilon}\right)\right) \\ &= \exp\left(-\boldsymbol{K}(\varepsilon;\theta)\right)\end{aligned}$$

where we note that the form given in precisely the relative entropy between Bernoulli(ε) and Bernoulli(θ).

Returning to the setting of interest with $x_1, \ldots, x_n$ and the sample mean $\bar{x}$, note that using Markov's inequality again and the iid assumption on the data, we have

$$\begin{aligned}\mathbf{P}\{\bar{x} > \theta + \varepsilon\} &= \mathbf{P}\left\{\sum_{i=1}^n x_i > n(\theta+\varepsilon)\right\} \\ &\leq \left(\exp(-t(\theta+\varepsilon))\, \mathbf{E}_\mu\, e^{tx}\right)^n.\end{aligned}$$

Setting $t = t^*(\varepsilon + \theta)$ then, and using a classical lower bound on the relative entropy (Lemma 16), we obtain

$$\begin{aligned}\mathbf{P}\{\bar{x} > \theta + \varepsilon\} &\leq \left(\exp\left(-\boldsymbol{K}(\theta+\varepsilon;\theta)\right)\right)^n \\ &\leq \left(\exp\left(-2((\theta+\varepsilon)-\theta)^2\right)\right)^n \\ &= \exp\left(-2n\varepsilon^2\right). \end{aligned} \tag{13}$$

Note that since $\varepsilon + \theta > \theta$ for all $\varepsilon > 0$, it follows that $t^*(\varepsilon + \theta) > 0$ for all $0 < \varepsilon < 1 - \theta$.

Next we seek a lower bound on $\bar{x} - \theta$, equivalently an upper bound on $-\bar{x} + \theta$. This can be done by essentially the same process. Again for $X \sim \text{Bernoulli}(\theta)$, using Markov's inequality, we have for any $s > 0$ that

$$\begin{aligned}\mathbf{P}\{X - \theta < -\varepsilon\} &= \mathbf{P}\{-X > \varepsilon - \theta\} \\ &= \mathbf{P}\{\exp(-sX) > \exp(s(\varepsilon-\theta))\} \\ &\leq \exp(-s(\varepsilon-\theta))\,\mathbf{E}\,e^{-sX} \\ &= \exp(s(\theta-\varepsilon))\left(1-\theta+\theta e^{-s}\right).\end{aligned}$$

This is, of course, a rather familiar form. Writing $a = \theta - \varepsilon$, note that the function

$$\exp(sa)\left(1-\theta+\theta e^{-s}\right)$$

is minimized as a function of s at

$$s^* = \log\left(\frac{1-a}{1-\theta}\right)\left(\frac{\theta}{a}\right),$$

which analogous to earlier in the proof, satisfies $s^* > 0$ only when $\theta > a = \theta - \varepsilon$, which is to say whenever $\varepsilon > 0$. Keeping with the a notation, note that plugging in s^* to the bound above, we have

$$\begin{aligned}\exp(s^*a)\left(1-\theta+\theta e^{-s^*}\right) &= \exp\left((1-a)\log\left(\frac{1-\theta}{1-a}\right) + a\log\left(\frac{\theta}{a}\right)\right) \\ &= \exp\left(-\boldsymbol{K}(a;\theta)\right),\end{aligned}$$

the exact same bound as before. It follows that

$$\begin{aligned}\mathbf{P}\{\bar{x} - \theta < -\varepsilon\} &= \mathbf{P}\{-\bar{x} > \varepsilon - \theta\} \\ &= \mathbf{P}\left\{-\sum_{i=1}^n x_i > n(\varepsilon-\theta)\right\} \\ &\leq \left(\exp(s(\theta-\varepsilon))\,\mathbf{E}_\mu\, e^{-sx}\right)^n.\end{aligned}$$

Setting $s = t^*$ with $a = \theta - \varepsilon$, in a form analogous to the upper bounds done earlier, we have

$$\begin{aligned}\mathbf{P}\{\bar{x} - \theta < -\varepsilon\} &\leq \left(\exp\left(-\boldsymbol{K}(\theta-\varepsilon;\theta)\right)\right)^n \\ &\leq \left(\exp\left(-2((\theta-\varepsilon)-\theta)^2\right)\right)^n \\ &= \exp\left(-2n\varepsilon^2\right).\end{aligned} \tag{14}$$

Taking a union bound over the two "bad events" in (13) and (14), we have

$$\begin{aligned}\mathbf{P}\{|\bar{x} - \theta| < -\varepsilon\} &\leq \mathbf{P}\{\bar{x} - \theta < -\varepsilon\} \cup \mathbf{P}\{\bar{x} - \theta > \varepsilon\} \\ &\leq \mathbf{P}\{\bar{x} - \theta < -\varepsilon\} + \mathbf{P}\{\bar{x} - \theta > \varepsilon\} \\ &\leq 2\exp\left(-2n\varepsilon^2\right),\end{aligned}$$

concluding the proof. □

Fundamental PAC-Bayes identity The following identity is fundamental to theoretical PAC-Bayesian analysis, and is a well-known result. Catoni [7, p. 159–160] for example gives a concise proof, but for completeness, we provide a step-by-step proof of this result here. The key elements of the following theorem are the prior $\nu \in \mathcal{M}_+^1$, and candidate posterior $\rho \in \mathcal{M}_+^1$.

Theorem 18. *For any measurable function h,*

$$\log \mathbf{E}_\nu \exp(h) = \sup_{\rho \in \mathcal{M}_+^1} \left(\sup_{b \in \mathbb{R}} \mathbf{E}_\rho (b \wedge h) - K(\rho; \nu) \right).$$

In the special case where h is bounded above, then the above equality simplifies to

$$\log \mathbf{E}_\nu \exp(h) = \sup_{\rho \in \mathcal{M}_+^1} \left(\mathbf{E}_\rho h - K(\rho; \nu) \right).$$

Proof of Theorem 18. The key to this proof is a simple expansion of the relative entropy between an arbitrary $\rho \in \mathcal{M}_+^1$ and a specially modified prior ν^*. This ν^* is defined in terms of the following requirement on the density function $d\nu^*/d\nu$: almost everywhere $[\nu]$, we must have

$$\left(\frac{d\nu^*}{d\nu} \right)(\omega) = g^*(\omega) \coloneqq \frac{\exp(h(\omega))}{\mathbf{E}_\nu \exp(h)}.$$

Satisfying this is easy by construction. Just define ν^* using g^*, as

$$\nu^*(A) \coloneqq \int_A g^* \, d\nu, \qquad A \in \mathcal{A}.$$

Since $g^* \geq 0$, it follows that ν^* is non-negative, and thus a measure on $(\Omega, \mathcal{A})$. As long as $\exp(h)$ is ν-integrable, we have

$$\int_A g^* \, d\nu = (\mathbf{E}_\nu \exp(h))^{-1} \int_A \exp(h(\omega)) \, d\nu(\omega) \leq 1,$$

and also that $\nu^*(\Omega) = 1$, so $\nu^* \in \mathcal{M}_+^1$. Furthermore, note that $\nu^* \ll \nu$ and $\nu \ll \nu^*$.

Now, before proving all the necessary facts, let us run through the primary step of the argument using the following series of identities, which should be rather intuitive even at first glance:

$$\begin{aligned}
\boldsymbol{K}(\rho; \nu^*) &= \mathbf{E}_\rho \log \left(\frac{d\rho}{d\nu^*} \right) \\
&= \mathbf{E}_\rho \log \left(\frac{d\rho}{d\nu} \frac{d\nu}{d\nu^*} \right) && (15) \\
&= \mathbf{E}_\rho \left(\log \frac{d\rho}{d\nu} + \log \frac{d\nu}{d\nu^*} \right) && (16) \\
&= \mathbf{E}_\rho \left(\log \frac{d\rho}{d\nu} + \log \mathbf{E}_\nu \exp(h) - h \right).
\end{aligned}$$

When the left-hand side is finite, so is the right-hand side, and they are equal. Furthermore, when the left-hand side is infinite, so is the right-hand side.

To prove the above chain of equalities, first start by writing $g(\omega) = (d\nu^*/d\nu)(\omega)$, and observe that by the chain rule (Lemma 14), we have

$$\int_A \left(\frac{1}{g(\omega)} \right) d\nu^* = \int_A \left(\frac{1}{g(\omega)} \right) g(\omega) \, d\nu(\omega) = \nu(A),$$

for any $A \in \mathcal{A}$. By $\nu \ll \nu^*$ and the Radon-Nikodym theorem, it follows that almost everywhere $[\nu]$, we have

$$\left(\frac{d\nu}{d\nu^*} \right)(\omega) = \frac{1}{g(\omega)} = \frac{\mathbf{E}_\nu \exp(h)}{\exp(h(\omega))}, \tag{17}$$

which justifies writing $d\nu/d\nu^* = 1/(d\nu^*/d\nu)$. Another basic fact using the chain rule (Lemma 14) is that for each $A \in \mathcal{A}$,

$$\begin{aligned}
\int_A \left(\frac{d\rho}{d\nu} \right)(\omega) \left(\frac{\mathbf{E}_\nu \exp(h)}{\exp(h(\omega))} \right) d\nu^*(\omega) &= \int_A \left(\frac{d\rho}{d\nu} \right)(\omega) \left(\frac{\mathbf{E}_\nu \exp(h)}{\exp(h(\omega))} \right) \left(\frac{\exp(h(\omega))}{\mathbf{E}_\nu \exp(h)} \right) d\nu(\omega) \\
&= \int_A \frac{d\rho}{d\nu} \, d\nu \\
&= \rho(A) \\
&= \int_A \frac{d\rho}{d\nu^*} \, d\nu^*
\end{aligned}$$

where the final three equalities follow from the Radon-Nikodym theorem and $\rho \ll \nu$ and $\rho \ll \nu^*$. Taking this basic fact and plugging in (17), we have

$$\int_A \frac{d\rho}{d\nu}\frac{d\nu}{d\nu^*}\,d\nu^* = \int_A \frac{d\rho}{d\nu^*}\,d\nu^*, \qquad A \in \mathcal{A}$$

and then by uniqueness of the density function, that almost everywhere $[\nu^*]$,

$$\frac{d\rho}{d\nu}\frac{d\nu}{d\nu^*} = \frac{d\rho}{d\nu^*}.$$

Since any statement a.e. $[\nu^*]$ holds a.e. $[\rho]$ by $\rho \ll \nu^*$, this proves (15).

The first equality holds from the definition of relative entropy, and with (15) now established, the remaining two equalities follow immediately from (17).

The next step is to show that we can meaningfully write

$$\mathbf{E}_\rho\left(\log\frac{d\rho}{d\nu} + \log\mathbf{E}_\nu\exp(h) - h\right) = \boldsymbol{K}(\rho;\nu) + \log\mathbf{E}_\nu\exp(h) - \mathbf{E}_\rho\, h \tag{18}$$

in the sense that both sides are well-defined, and take on equal values in $\mathbb{R}\cup\{\infty\}$. To prove this, we would like to use the basic additivity property of Lebesgue integrals [4, Theorem 1.6.3]. First observe that the integrand of the left-hand side is well-defined and equal to $\boldsymbol{K}(\rho;\nu^*)$. We need to show that the right-hand side is also well-defined. The first term $\boldsymbol{K}(\rho;\nu) \geq 0 > -\infty$ by Lemma 15, and thus while it cannot be $-\infty$, it takes values in $\mathbb{R}\cup\{\infty\}$. The remaining term depends on h. In the case that h is bounded above, we have that $\mathbf{E}_\rho\, h < \infty$, meaning that the right-hand side of (18) is well-defined, which implies via additivity that both sides of (18) take values in $\mathbb{R}\cup\{\infty\}$, and are equal in both the finite and infinite cases.

Note that when h is not bounded above, this leaves the possibility that $\mathbf{E}_\rho\, h = \infty$, which would lead to the ambiguous $\infty - \infty$ on the right-hand side of (18), spoiling the additivity property.

With the assumption of h bounded above, and re-arranging some terms, we can write

$$\boldsymbol{K}(\rho;\nu^*) = \log\mathbf{E}_\nu\exp(h) - \left(\mathbf{E}_\rho\, h - \boldsymbol{K}(\rho;\nu)\right).$$

By non-negativity of the relative entropy (Lemma 15), the left-hand side is minimized when $\rho = \nu^*$, in which case it takes the value $\boldsymbol{K}(\nu^*;\nu^*) = 0$. Note that as $\nu^* \in \mathcal{M}_+^1$, the supremum of the term in parentheses on the right-hand side is achieved at $\rho = \nu^*$. This means we can write

$$\log\mathbf{E}_\nu\exp(h) = \sup_{\rho\in\mathcal{M}_+^1}\left(\mathbf{E}_\rho\, h - \boldsymbol{K}(\rho;\nu)\right) \tag{19}$$

for h bounded above.

To complete the proof, we must consider the case where h is unbounded. As preparation, create a measurable function sequence (h_k) defined by $h_k = b_k \wedge h$, where (b_k) satisfies $b_k \uparrow \infty$ and is increasing. Since we have

$$\lim_{k\to\infty}\exp(h_k(\omega)) = \exp(h(\omega))$$

pointwise in $\omega\in\Omega$, and $h_k \leq h_{k+1} \leq \ldots \leq h$ for any k, by the monotone convergence theorem, we have

$$\lim_{k\to\infty}\mathbf{E}_\nu\exp(h_k) = \mathbf{E}_\nu\exp(h),$$

and using the continuity of the log function,

$$\lim_{k\to\infty}\log\mathbf{E}_\nu\exp(h_k) = \log\left(\lim_{k\to\infty}\mathbf{E}_\nu\exp(h_k)\right) = \log\mathbf{E}_\nu\exp(h).$$

This means we can write

$$\begin{aligned}
\log\mathbf{E}_\nu\exp(h) &= \sup_{b\in\mathbb{R}}\log\mathbf{E}_\nu\exp(b\wedge h)\\
&= \sup_{b\in\mathbb{R}}\sup_{\rho\in\mathcal{M}_+^1}\left(\mathbf{E}_\rho(b\wedge h) - \boldsymbol{K}(\rho;\nu)\right) &(20)\\
&= \sup_{\rho\in\mathcal{M}_+^1}\sup_{b\in\mathbb{R}}\left(\mathbf{E}_\rho(b\wedge h) - \boldsymbol{K}(\rho;\nu)\right) &(21)\\
&= \sup_{\rho\in\mathcal{M}_+^1}\left(\sup_{b\in\mathbb{R}}\mathbf{E}_\rho(b\wedge h) - \boldsymbol{K}(\rho;\nu)\right).
\end{aligned}$$

Since for any $b \in \mathbb{R}$, we have that $b \wedge h \leq b < \infty$, we can use (19), the key identity for the case of bounded functions, which immediately implies (20). Finally, regarding the swap of supremum operations, note that the function of interest is

$$f(\rho, b) = \mathbf{E}_\rho(b \wedge h) - \boldsymbol{K}(\rho; \nu), \qquad (\rho, b) \in \mathcal{M}_+^1 \times \mathbb{R}.$$

For an arbitrary sequence (ρ_k, b_k), observe that for all k,

$$f(\rho_k, b_k) \leq \sup_\rho f(\rho, b_k) \leq \sup_b \sup_\rho f(\rho, b)$$
$$f(\rho_k, b_k) \leq \sup_b f(\rho_k, b) \leq \sup_\rho \sup_b f(\rho, b).$$

If f is unbounded on $\mathcal{M}_+^1 \times \mathbb{R}$, then the sequence (ρ_k, b_k) can be constructed such that $f(\rho_k, b_k) \to \infty$ as $k \to \infty$, implying that in both cases the supremum is infinite, so equality holds trivially. On the other hand, when f is bounded above, the sequence can be constructed such that $f(\rho_k, b_k) \to B$, and so the above inequalities imply

$$B = \lim_{k\to\infty} f(\rho_k, b_k) \leq \sup_b \sup_\rho f(\rho, b) \leq B$$
$$B = \lim_{k\to\infty} f(\rho_k, b_k) \leq \sup_\rho \sup_b f(\rho, b) \leq B$$

and thus, as desired, the step to (21) holds. This concludes the chain of equalities and the proof. □

A.2 Proofs of results in the main text

Proof of Lemma 3. Start with the following elementary inequality: if X is a random variable such that $\mathbf{E}\, e^X \leq 1$, then for any $\delta \in (0, 1)$, we have that X exceeds $\log(\delta^{-1})$ with probability no greater than δ. To see this, observe that

$$\mathbf{P}\{X \geq \log(\delta^{-1})\} = \mathbf{P}\{\exp(X) \geq 1/\delta\} = \mathbf{E}\, I\{\delta \exp(X) \geq 1\} \leq \mathbf{E}\, \delta e^X \leq \delta. \tag{22}$$

Next, we set the function h in Theorem 18 to be a sum of functions depending on both the data and the noise, as

$$h(\epsilon) = \sum_{i=1}^n f(x_i, \epsilon) - n \log \mathbf{E}_\mu \exp(f(x, \epsilon)).$$

Since f is bounded on $\mathbb{R}^2$ by hypothesis, we have that h is also bounded. Using Theorem 18, we have

$$\begin{aligned} B_0 &\coloneqq \sup_{\rho \in \mathcal{M}_+^1} (\mathbf{E}_\rho\, h(\epsilon) - K(\rho; \nu)) \\ &= \log \mathbf{E}_\nu \left(\frac{\exp\left(\sum_{i=1}^n f(x_i, \epsilon)\right)}{(\mathbf{E}_\mu \exp(f(x, \epsilon)))^n} \right). \end{aligned}$$

Next, taking expectation with respect to the sample, observe that

$$\begin{aligned} \mathbf{E} \exp(B_0) &= \mathbf{E} \int \left(\frac{\exp\left(\sum_{i=1}^n f(x_i, \epsilon)\right)}{(\mathbf{E}_\mu \exp(f(x, \epsilon)))^n} \right) \nu(\epsilon) \\ &= \int \left(\frac{\mathbf{E} \exp\left(\sum_{i=1}^n f(x_i, \epsilon)\right)}{(\mathbf{E}_\mu \exp(f(x, \epsilon)))^n} \right) \nu(\epsilon) \\ &= 1. \end{aligned}$$

The above equalities follow from straightforward algebraic manipulations, independence of the data, and taking the integration over the sample inside the integration over the noise, valid using Fubini's theorem. Applying (22) with $X = B_0$, noting that the only randomness is due to the sample, it holds that for probability at least $1 - \delta$, uniform in the choice of ρ, we have

$$\mathbf{E}_\rho\, h(\epsilon) - K(\rho; \nu) \leq \log(\delta^{-1}).$$

Plugging in the above definition of h and dividing by n, we have

$$\frac{1}{n} \sum_{i=1}^n \int f(x_i, \epsilon)\, d\rho(\epsilon) \leq \int \log \mathbf{E}_\mu \exp(f(x, \epsilon))\, d\rho(\epsilon) + \frac{\boldsymbol{K}(\rho; \nu) + \log(\delta^{-1})}{n}.$$

Finally, since the noise observations are iid, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\int f(x_i,\epsilon)\,d\rho(\epsilon) &= \frac{1}{n}\sum_{i=1}^{n}\int f(x_i,\epsilon_i)\,d\rho(\epsilon_i)\\ &= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i,\epsilon_i)\right)\end{aligned}$$

with expectation over the noise sample. This equality yields the desired result. □

Proof of Proposition 4. First, note that the upper bound derived from (5) holds uniformly in the choice of θ on a $(1-\delta)$ high-probability event. Setting $\theta = 1/2$ and solving for the optimal $s > 0$ setting is just calculus. It remains to obtain a corresponding lower bound on $\widehat{x} - \mathbf{E}_\mu\, x$. To do so, consider the analogous setting of Bernoulli ν and ρ, but this time on the domain $\{-1, 0\}$, with $\rho\{-1\} = \theta$ and $\nu\{-1\} = 1/2$. Using (1) and Lemma 3 again, we have

$$\left(\frac{-\theta}{s}\right)\widehat{x} \le \frac{-\theta\,\mathbf{E}_\mu\, x}{s} + \frac{\theta\,\mathbf{E}_\mu\, x^2}{2s^2} + \frac{1}{n}\left(\theta\log(2\theta) + (1-\theta)\log(2(1-\theta)) + \log(\delta^{-1})\right)$$

where we note $\mathbf{E}_\rho\,\epsilon = -\theta$ and $\mathbf{E}_\rho\,\epsilon^2 = \mathbf{E}_\rho\,|\epsilon| = \theta$. This yields a high-probability lower bound in the desired form when we set $\theta = 1/2$, since an upper bound on $-\widehat{x} + \mathbf{E}_\mu\, x$ is equivalent to a lower bound on $\widehat{x} - \mathbf{E}_\mu\, x$. However, since we have changed the prior in this case, the high-probability event here need not be the same as that for the upper bound, and as such, we must take a union bound over these two events to obtain the desired final result. □

Proof of Theorem 6. For clean notation, denote the empirical risk as

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} l(h;\boldsymbol{z}_i), \qquad h \in \mathcal{H}.$$

Using a classical Chernoff bound specialized to the case of Bernoulli observations (Lemma 17), we have that for any $h \in \mathcal{H}$, it holds that

$$\mathbf{P}\left\{R(h) - \widehat{R}(h) > \varepsilon\right\} \le \exp\left(-2n\varepsilon^2\right).$$

Rearranging terms, it follows immediately that with probability no less than $1 - \nu(h)\,\delta$, we have

$$R(h) - \widehat{R}(h) \le \varepsilon^*(h) \coloneqq \sqrt{\frac{\log(1/\nu(h)) + \log(1/\delta)}{2n}}.$$

The desired result follows from a union bound:

$$\begin{aligned}\mathbf{P}\left\{\exists\, h \in \mathcal{H} \text{ s.t. } R(h) - \widehat{R}(h) > \varepsilon^*(h)\right\} &\le \mathbf{P}\bigcup_{h\in\mathcal{H}}\left\{R(h) - \widehat{R}(h) > \varepsilon^*(h)\right\}\\ &\le \sum_{h\in\mathcal{H}}\mathbf{P}\left\{R(h) - \widehat{R}(h) > \varepsilon^*(h)\right\}\\ &\le \sum_{h\in\mathcal{H}}\nu(h)\delta\\ &= \delta.\end{aligned}$$

The event on the left-hand side of the above inequality is precisely that of the hypothesis, namely the "bad event" on which the sample is such that the risk $R(h)$ exceeds the given bound for *some* candidate $h \in \mathcal{H}$. □

Proof of Theorem 7. We start by making use of the pointwise deviation bound given in Proposition 4, which tells us that with $(1-2\delta)$ high probability

$$R(h) \le \frac{s}{n}\sum_{i=1}^{n}\psi\left(\frac{l(h;\boldsymbol{z}_i)}{s}\right) + \sqrt{\frac{2m_2(h)\log(\delta^{-1})}{n}}$$

for any pre-fixed $h \in \mathcal{H}$. Replacing δ with $\nu(h)\delta$ gives the key error level

$$\varepsilon^*(h) \coloneqq \sqrt{\frac{2m_2(h)\left(\log(1/\nu(h)) + \log(1/\delta)\right)}{n}},$$

and using the union bound argument in the proof of Theorem 6, we have

$$\mathbf{P}\left\{\exists\, h \in \mathcal{H} \text{ s.t. } R(h) - \widehat{R}_\psi(h) > \varepsilon^*(h)\right\} \leq 2\delta.$$

□

Proof of Theorem 9. To begin, let us recall a useful "change of measures" inequality,[6] which can be immediately derived from our proof of Theorem 18. In particular, recall from identity (18) that given some prior ν and constructing ν^* such that almost everywhere $[\nu]$ one has

$$\left(\frac{d\nu^*}{d\rho}\right)(h) = \frac{\exp(\varphi(h))}{\mathbf{E}_\nu \exp(\varphi)},$$

it follows that

$$\begin{aligned}\boldsymbol{K}(\rho;\nu^*) &= \mathbf{E}_\rho\left(\log\frac{d\rho}{d\nu} + \log\mathbf{E}_\nu\exp(\varphi) - \varphi\right)\\ &= \boldsymbol{K}(\rho;\nu) + \log\mathbf{E}_\nu\exp(\varphi) - \mathbf{E}_\rho\,\varphi\end{aligned}$$

whenever $\mathbf{E}_\rho\,\varphi < \infty$. In the case where $\mathbf{E}_\rho\,\varphi = \infty$, upper bounds are of course meaningless. Re-arranging, observe that since $\boldsymbol{K}(\rho;\nu^*) \geq 0$, it follows that

$$\mathbf{E}_\rho\,\varphi \leq \boldsymbol{K}(\rho;\nu) + \log\mathbf{E}_\nu\exp(\varphi). \tag{23}$$

This inequality given in (23) is deterministic, holds for any choice of ρ, and is a standard technical tool in deriving PAC-Bayes bounds.

We shall introduce a minor modification to this now-standard strategy in order to make the subsequent results more lucid. Instead of ν^* as just characterized above, define ν_n^* such that almost surely $[\nu]$, we have

$$\left(\frac{d\nu_n^*}{d\rho}\right)(h) = g(h) \coloneqq \frac{\exp(\varphi(h))}{\mathbf{E}_\nu\exp(\varphi/c_n)},$$

where $1 \leq c_n < \infty$ is a function of the sample size n, which increases monotonically as $c_n \uparrow \infty$ when $n \to \infty$ (e.g., setting $c_n = \sqrt{n}$). To explicitly construct such a measure, one can define it by $\nu_n^*(A) \coloneqq \int_A g\,d\nu$, for all $A \subset \mathcal{A}$, where $(\mathcal{H},\mathcal{A})$ is our measurable space of interest. In this paper, we always[7] have $\varphi > -\infty$, implying that $\mathbf{E}_\nu\exp(\varphi) > 0$. Also by assumption, since R is bounded over $h \in \mathcal{H}$, we have $\mathbf{E}_\nu\exp(\varphi) < \infty$, which in turn implies

$$0 < \nu_n^*(\mathcal{H}) = \frac{\mathbf{E}_\nu\exp(\varphi)}{\mathbf{E}_\nu\exp(\varphi/c_n)} < \infty,$$

and so ν_n^* is a finite measure. Note however that both $\nu_n^*(\mathcal{H}) > 1$ and $\nu_n^*(\mathcal{H}) < 1$ are possible, so in general ν_n^* need not be a probability measure. By construction, we have $\nu_n^* \ll \nu$. Since $\varphi(h) > -\infty$ for all $h \in \mathcal{H}$, we have that $g > 0$ and thus the measurability of g implies the measurability of $1/g$. Using the chain rule (Lemma 14), it follows that for any $A \in \mathcal{A}$,

$$\int_A \left(\frac{1}{g}\right) d\nu_n^* = \int_A \left(\frac{1}{g}\right)(g)\; d\nu = \nu(A).$$

As such, we have $\nu \ll \nu_n^*$, and by the Radon-Nikonym theorem, we may write $1/g = d\nu/d\nu_n^*$ since such a function is unique almost everywhere $[\nu_n^*]$. As long as $\rho \ll \nu$, which in turn implies $\rho \ll \nu_n^*$, so that with use of the chain rule and Radon-Nikodym, we have

$$\int_A \left(\frac{d\rho}{d\nu}\right)\left(\frac{1}{g}\right) d\nu_n^* = \int_A \left(\frac{d\rho}{d\nu}\right)\left(\frac{1}{g}\right) g\,d\nu = \rho(A) = \int_A \left(\frac{d\rho}{d\nu_n^*}\right) d\nu_n^*.$$

[6] There are other very closely related approaches to this proof. See Tolstikhin and Seldin [21], Bégin et al. [5] for some recent examples. Furthermore, we note that the key facts used here are also present in Catoni [8].

[7] We will only be using $\varphi \propto R - \widehat{R}_\psi$, so this statement holds via $R \geq 0$ and $\|R_\psi\|_\infty < \infty$.

Taking the two ends of this string of equalities, by Radon-Nikodym it holds that

$$\frac{d\rho}{d\nu}\frac{d\nu}{\nu_n^*} = \frac{d\rho}{d\nu_n^*}$$

a.e. $[\nu_n^*]$, and thus a.e. $[\rho]$ as well. Following the argument of Theorem 18, we have that

$$\boldsymbol{K}(\rho;\nu_n^*) = \boldsymbol{K}(\rho;\nu) + \log \mathbf{E}_\nu \exp(\varphi/c_n) - \mathbf{E}_\rho\, \varphi.$$

The tradeoff for using ν_n^* which need not be a probability comes in deriving a lower bound on $\boldsymbol{K}(\nu;\nu_n^*)$. In Lemma 15 we showed how the relative entropy between probability measures is non-negative. Non-negativity does not necessarily hold for general measures, but analogous lower bounds can be readily derived for our special case as

$$\boldsymbol{K}(\rho;\nu_n^*) = \mathbf{E}_\rho \log \frac{d\rho}{d\nu_n^*} = \mathbf{E}_{\nu_n^*} \frac{d\rho}{d\nu_n^*} \log \frac{d\rho}{d\nu_n^*} \geq \mathbf{E}_{\nu_n^*} \left(\frac{d\rho}{d\nu_n^*} - 1 \right) = 1 - \nu_n^*(\mathcal{H}),$$

where the last inequality uses the fact that ρ is a probability and $\rho(A) = \int_A (d\rho/d\nu_n^*)\, d\nu_n^*$ for all $A \in \mathcal{A}$. Taking this with our decomposition of $\boldsymbol{K}(\rho;\nu_n^*)$, we have

$$\mathbf{E}_\rho\, \varphi \leq \boldsymbol{K}(\rho;\nu) + \log \mathbf{E}_\nu \exp\left(\varphi/c_n\right) - 1 + \nu_n^*(\mathcal{H}), \tag{24}$$

which amounts to a revised inequality based on change of measures, analogous to (23).

To keep notation clean, write

$$\begin{aligned} X(h) &\coloneqq R(h) - \frac{s}{n}\sum_{i=1}^n \psi\left(\frac{l(h;\boldsymbol{z}_i)}{s}\right) = R(h) - \widehat{R}_\psi(h) \\ m_2(h) &\coloneqq \mathbf{E}_\mu\, l(h;\boldsymbol{z})^2 \\ v(h) &\coloneqq \mathbf{E}_\mu (l(h;\boldsymbol{z}) - R(h))^2 \end{aligned}$$

Noting that $X(h)$ is random with dependence on the sample, via Markov's inequality we have

$$\mathbf{E}_\nu\, e^X \leq \frac{\mathbf{E}_n\, \mathbf{E}_\nu\, e^X}{\delta}, \tag{25}$$

with probability no less than $1-\delta$. Here probability and $\mathbf{E}_n$ are with respect to the sample. Since $\widehat{R}_\psi$ is bounded, as long as $\mathbf{E}_\rho\, R < \infty$, we have $\mathbf{E}_\rho\, X < \infty$, which lets us use the change of measures inequality in a meaningful way. Now for $c_n > 0$, observe that we have

$$\begin{aligned} c_n\, \mathbf{E}_\rho\, X = \mathbf{E}_\rho\, c_n X &\leq \boldsymbol{K}(\rho;\nu) + \log \mathbf{E}_\nu \exp\left(X\right) - 1 + \nu_n^*(\mathcal{H}) \\ &\leq \boldsymbol{K}(\rho;\nu) + \log(\delta^{-1}) + \log \mathbf{E}_n\, \mathbf{E}_\nu \exp\left(X\right) - 1 + \nu_n^*(\mathcal{H}) \\ &= \boldsymbol{K}(\rho;\nu) + \log(\delta^{-1}) + \log \mathbf{E}_\nu\, \mathbf{E}_n \exp\left(X\right) - 1 + \nu_n^*(\mathcal{H}) \end{aligned}$$

with probability no less than $1-\delta$. The first inequality follows from modified change of measures (24), the second inequality follows from (25), and the final interchange of integration operations is valid using Fubini's theorem [4]. Note that the $1-\delta$ "good event" depends only on ν (fixed in advance) and not ρ. Thus, the above inequality holds on the good event, uniformly in ρ.

It remains to bound $\mathbf{E}_n \exp(cX)$, for an arbitrary constant $c > 0$ (here we will have $c = 1$). Start by breaking up the one-sided deviations as

$$X = R - \widehat{R}_\psi = \left(R - \mathbf{E}_n\, \widehat{R}_\psi\right) + \left(\mathbf{E}_n\, \widehat{R}_\psi - \widehat{R}_\psi\right),$$

writing $X_{(1)} \coloneqq R - \mathbf{E}_n\, \widehat{R}_\psi$ and $X_{(2)} \coloneqq \mathbf{E}_n\, \widehat{R}_\psi - \widehat{R}_\psi$ for convenience. We will take the terms $X_{(1)}$ and $X_{(2)}$ one at a time. First, note that the function ψ can be written

$$\psi(u) = \left(u - \frac{u^3}{6}\right)\left(I\{u \leq \sqrt{2}\} - I\{u < -\sqrt{2}\}\right) + \frac{2\sqrt{2}}{3}\left(1 - I\{u \leq \sqrt{2}\} - I\{u < -\sqrt{2}\}\right). \tag{26}$$

Again for notational simplicity, write $L = l(h;\boldsymbol{z})$ and $L_i = l(h;\boldsymbol{z}_i)$, $i \in [n]$, where $h \in \mathcal{H}$ is arbitrary. Write $\mathcal{E}_i^+ \coloneqq \left\{L_i \leq s\sqrt{2}\right\}$ and $\mathcal{E}_i^- \coloneqq \left\{L_i < -s\sqrt{2}\right\}$. We are assuming non-negative losses, so that

$L \geq 0$. This means that $I(\mathcal{E}_i^-) = 0$ and $\mathbf{P}\,\mathcal{E}_i^- = 0$. We use this, as well as $1 - \mathbf{P}\,\mathcal{E}_i^+ \geq 0$, in addition to (26) in order to bound the expectation of our estimator $\widehat{R}_\psi$ from below, as follows.

$$\begin{aligned}
\mathbf{E}_n\,\widehat{R}_\psi &= \frac{s}{n}\sum_{i=1}^{n}\mathbf{E}_\mu\,\psi\left(\frac{L_i}{s}\right) \\
&= \frac{s}{n}\sum_{i=1}^{n}\left[\mathbf{E}_\mu\left(\frac{L_i}{s} - \frac{L_i^3}{6s^3}\right)\left(I(\mathcal{E}_i^+) - I(\mathcal{E}_i^-)\right) + \frac{2\sqrt{2}}{3}\left(1 - \mathbf{P}\,\mathcal{E}_i^+ - \mathbf{P}\,\mathcal{E}_i^-\right)\right] \\
&\geq \frac{s}{n}\sum_{i=1}^{n}\mathbf{E}_\mu\left(\frac{L_i}{s} - \frac{L_i^3}{6s^3}\right) I(\mathcal{E}_i^+) \\
&= \mathbf{E}_\mu\,L\,I\{L \leq s\sqrt{2}\} - \frac{1}{6s^2}\,\mathbf{E}_\mu\,L^3 I\{L \leq s\sqrt{2}\} \\
&= R - \mathbf{E}_\mu\,L\,I\{X > s\sqrt{2}\} - \frac{1}{6s^2}\,\mathbf{E}_\mu\,L^3 I\{L \leq s\sqrt{2}\}.
\end{aligned}$$

By assumption, we have $\mathbf{E}_\mu\,L^3 I\{L \leq s\sqrt{2}\} \leq \mathbf{E}\,L^3 \leq M_3 < \infty$, implying that this lower bound is non-trivial. Next we obtain a one-sided bound on the tails of the loss by

$$\begin{aligned}
\mathbf{P}\left\{L > s\sqrt{2}\right\} &= \mathbf{P}\left\{L - R > s\sqrt{2} - R\right\} \\
&\leq \mathbf{P}\left\{|L - R| > s\sqrt{2} - R\right\} \\
&\leq \frac{\mathbf{E}_\mu\,|L - R|^2}{\left(s\sqrt{2} - R\right)^2}.
\end{aligned}$$

Note that the first inequality makes use of $s\sqrt{2} > R$, which is implied by the bounds assumed on R, namely that $1/2 \geq R\sqrt{\log(\delta^{-1})/(nM_2)}$.

Returning to the lower bound on $\widehat{R}_\psi$, using Hölder's inequality in conjunction with the tail bound we just obtained, we get an upper bound in the form of

$$\begin{aligned}
\mathbf{E}_\mu\,L\,I\{L > s\sqrt{2}\} &= \mathbf{E}_\mu\,|L\,I\{L > s\sqrt{2}\}| \\
&\leq \sqrt{\mathbf{E}_\mu\,L^2\,\mathbf{P}\{L > s\sqrt{2}\}} \\
&\leq \sqrt{\frac{\mathbf{E}_\mu\,L^2\,\mathbf{E}_\mu\,|L - R|^2}{\left(s\sqrt{2} - R\right)^2}}.
\end{aligned}$$

This means we can now say

$$\mathbf{E}_n\,\widehat{R}_\psi \geq R - \sqrt{\frac{\mathbf{E}_\mu\,L^2\,\mathbf{E}_\mu\,|L - R|^2}{\left(s\sqrt{2} - R\right)^2}} - \frac{1}{6s^2}\,\mathbf{E}_\mu\,L^3 I\{L \leq s\sqrt{2}\},$$

which re-arranged and written more succinctly gives us

$$\begin{aligned}
X_{(1)}(h) &\leq \sqrt{\frac{m_2(h)v(h)}{\left(s\sqrt{2} - R\right)^2}} + \frac{\mathbf{E}_\mu\,|l(h;\boldsymbol{z})|^3}{6s^2} \\
&\leq \sqrt{\frac{M_2 V}{\left(s\sqrt{2} - R\right)^2}} + \frac{M_3}{6s^2} \\
&= \sqrt{\frac{V\log(\delta^{-1})}{n\left(1 - R\sqrt{\log(\delta^{-1})/(nM_2)}\right)^2}} + \frac{M_3\log(\delta^{-1})}{3M_2 n} \\
&\leq B_{(1)} \\
&:= 2\sqrt{\frac{V\log(\delta^{-1})}{n}} + \frac{M_3\log(\delta^{-1})}{3M_2 n}
\end{aligned}$$

as desired. The final inequality uses the assumed bound on R. Note that this is a deterministic bound, in that it is free of both the choice of h (i.e., random draw from ν or ρ) and the sample, which we are integrating over.

Next, we look at the remaining deviations $X_{(2)} = \mathbf{E}_n \widehat{R}_\psi - \widehat{R}_\psi$. Writing $Y_i \coloneqq (s/n)\psi(L_i/s)$, we have $X_{(2)} = \sum_{i=1}^n (\mathbf{E}\, Y_i - Y_i)$. Since $0 \leq \psi(u) \leq 2\sqrt{2}/3$ for $u \geq 0$, and $L \geq 0$, we have that $0 \leq Y_i \leq 2\sqrt{2}s/(3n)$. It follows from Hoeffding's inequality that for all $\epsilon > 0$, we have

$$\begin{aligned} \mathbf{P}\left\{X_{(2)} > \epsilon\right\} &\leq \exp\left(\frac{-2\epsilon^2}{n(2\sqrt{2}s/(3n))^2}\right) \\ &= \exp\left(\frac{-9\epsilon^2 \log(\delta^{-1})}{2M_2}\right). \end{aligned} \tag{27}$$

Note that this bound does not depend on the setting of $\delta \in (0,1)$, which is fixed in advance. Also note that while we are dealing with the sum of bounded, independent random variables, the scaling factor $s \propto \sqrt{n}$ makes it such that these deviations converge to some potentially non-zero constant in the $n \to \infty$ limit, which is why n does not appear in the exponential on the right-hand side.

In any case, we can still readily use these sub-Gaussian tail bounds to control the expectation. Using the classic identity relating the expectation to the tails of a distribution,

$$\begin{aligned} \mathbf{E}_n \exp\left(cX_{(2)}\right) &= \int_0^\infty \mathbf{P}\left\{\exp\left(cX_{(2)}\right) > \epsilon\right\}\, d\epsilon \\ &= \int_{-\infty}^\infty \mathbf{P}\left\{\exp\left(cX_{(2)}\right) > \exp(\epsilon)\right\} \exp(\epsilon)\, d\epsilon \end{aligned} \tag{28}$$

where the second equality follows using integration by substitution. The right-hand side of (28) is readily controlled as follows. Using (27) above, we have

$$\mathbf{P}\left\{\exp\left(cX_{(2)}\right) > \exp(\epsilon)\right\} = \mathbf{P}\left\{X_{(2)} > \epsilon/c\right\} \leq \exp\left(\frac{-\epsilon^2}{2\sigma^2}\right)$$

where we have set $\sigma^2 \coloneqq c^2 M_2/(9\log(\delta^{-1}))$. The key bound of interest can be compactly written as

$$\begin{aligned} \mathbf{E}_n \exp\left(cX_{(2)}\right) &\leq 2\int_{-\infty}^\infty \exp\left(-\frac{\epsilon^2}{2\sigma^2} + \epsilon\right)\, d\epsilon \\ &= 2\int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}\left(\epsilon - \sigma^2\right)^2 + \frac{\sigma^2}{2}\right)\, d\epsilon \\ &= 2\sqrt{2\pi}\sigma \exp\left(\frac{\sigma^2}{2}\right) \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\epsilon - \sigma^2\right)^2\right)\, d\epsilon \\ &= 2\sqrt{2\pi}\sigma \exp\left(\frac{\sigma^2}{2}\right). \end{aligned}$$

Note that the first equality uses the usual "complete the square" identity, and the rest follows from basic properties of the Gaussian integral. Filling in the definition of σ, we have

$$\mathbf{E}_n \exp\left(cX_{(2)}\right) \leq 2\sqrt{2\pi}\left(c\sqrt{\frac{M_2}{9\log(\delta^{-1})}}\right) \exp\left(\frac{c^2 M_2}{9\log(\delta^{-1})}\right).$$

The right-hand side of this inequality is free of the choice of $h \in \mathcal{H}$, and thus taking expectation with respect to ν yields the same bound, i.e., the same bound holds for $\mathbf{E}_\nu\, \mathbf{E}_n \exp(cX_{(2)})$. Taking the log of this upper bound, we thus may conclude that

$$\begin{aligned} \log \mathbf{E}_\nu\, \mathbf{E}_n \exp(cX_{(2)}) &\leq \frac{1}{2}\left[\log(8\pi M_2 c^2) - \log(9\log(\delta^{-1}))\right] + \frac{c^2 M_2}{9\log(\delta^{-1})} \\ &\leq \frac{1}{2}\log(8\pi M_2 c^2) + c^2 M_2 \end{aligned}$$

on an event of probability no less than $1-\delta$. The latter inequality uses $\delta \leq \exp(-1/9)$. For the result of interest here, we can let $c = 1$.

Finally, going back to the bound on $c_n \, \mathbf{E}_\rho \, X$, we can control the key term by

$$\begin{aligned}
\log \mathbf{E}_\nu \, \mathbf{E}_n \exp\left(X\right) &= \log \mathbf{E}_\nu \, \mathbf{E}_n \exp\left(X_{(1)} + X_{(2)}\right) \\
&= \log \mathbf{E}_\nu \left[\exp\left(X_{(1)}\right) \mathbf{E}_n \exp\left(X_{(2)}\right)\right] \\
&\leq B_{(1)} + \log \mathbf{E}_\nu \, \mathbf{E}_n \exp\left(X_{(2)}\right) \\
&\leq B_{(1)} + \frac{1}{2} \log(8\pi M_2 c^2) + c^2 M_2.
\end{aligned}$$

Setting $c_n = \sqrt{n}$, we have

$$\begin{aligned}
\sqrt{n} \, \mathbf{E}_\rho \, X &\leq \boldsymbol{K}(\rho; \nu) + \log(\delta^{-1}) + \log \mathbf{E}_\nu \, \mathbf{E}_n \exp\left(X\right) - 1 + \nu_n^*(\mathcal{H}) \\
&\leq \boldsymbol{K}(\rho; \nu) + \log(\delta^{-1}) + 2\sqrt{\frac{V \log(\delta^{-1})}{n}} + \frac{M_3 \log(\delta^{-1})}{3 M_2 n} + \frac{\log(8\pi M_2)}{2} + M_2 - 1 + \nu_n^*(\mathcal{H}).
\end{aligned}$$

Dividing both sides by $\sqrt{n}$ yields the desired result. □

Proof of Proposition 12. To keep the notation clean, write $X = X(h) = -\sqrt{n}\widehat{R}_\psi(h)$. Similar to the proof of Theorem 18, we have

$$\begin{aligned}
\boldsymbol{K}(\rho; \widehat{\rho}) &= \mathbf{E}_\rho \log\left(\frac{d\rho}{d\widehat{\rho}}\right) \\
&= \mathbf{E}_\rho \log\left(\frac{d\rho}{d\nu}\frac{d\nu}{d\widehat{\rho}}\right) \\
&= \mathbf{E}_\rho \left(\log \frac{d\rho}{d\nu} + \log \mathbf{E}_\nu \exp(X) - X\right) \\
&= \boldsymbol{K}(\rho; \nu) + \log \mathbf{E}_\nu \exp(X) - \mathbf{E}_\rho \, X
\end{aligned}$$

whenever $\mathbf{E}_\rho \, X < \infty$. Using non-negativity of the relative entropy (Lemma 15), the left-hand side of this chain of equalities is minimized in ρ at $\rho = \widehat{\rho}$. Since $\log \mathbf{E}_\nu \exp(X)$ is free of ρ, it follows that

$$\begin{aligned}
\widehat{\rho} &\in \arg\min_\rho \left(\boldsymbol{K}(\rho; \nu) + \mathbf{E}_\rho(-1)X\right) \\
&= \arg\min_\rho \left(\frac{\boldsymbol{K}(\rho; \nu)}{\sqrt{n}} + \mathbf{E}_\rho \, \widehat{R}_\psi(h)\right) \\
&= \arg\min_\rho \left(\frac{\boldsymbol{K}(\rho; \nu)}{\sqrt{n}} + \mathbf{E}_\rho \, \widehat{R}_\psi(h) + C\right)
\end{aligned}$$

where C is any term which is constant in ρ, for example all the terms in the upper bound of Theorem 9 besides $\widehat{G}_{\rho,\psi} + \boldsymbol{K}(\rho; \nu)/\sqrt{n}$. This proves the result regarding the form of the new optimal Gibbs posterior.

Evaluating the risk bound under this posterior is straightforward computation. Observe that

$$\begin{aligned}
\boldsymbol{K}(\widehat{\rho}; \nu) &= \mathbf{E}_{\widehat{\rho}} \log \frac{d\widehat{\rho}}{d\nu} \\
&= \mathbf{E}_{\widehat{\rho}} \left(X(h) - \log \mathbf{E}_\nu \exp\left(X\right)\right) \\
&= -\sqrt{n} \, \mathbf{E}_{\widehat{\rho}} \, \widehat{R}_\psi - \log \mathbf{E}_\nu \exp\left(-\sqrt{n}\widehat{R}_\psi\right) \\
&= \log \mathbf{E}_\nu \exp\left(\sqrt{n}\widehat{R}_\psi\right) - \sqrt{n} \, \mathbf{E}_{\widehat{\rho}} \, \widehat{R}_\psi.
\end{aligned}$$

Substituting this into the upper bound of Theorem 9, the robust empirical mean estimate terms cancel, and we have

$$G_{\widehat{\rho}} := \mathbf{E}_{\widehat{\rho}} \, R \leq \frac{1}{\sqrt{n}} \left(\log \mathbf{E}_\nu \exp\left(\sqrt{n}\widehat{R}_\psi\right) + \frac{\log(8\pi M_2 \delta^{-2})}{2} + M_2 + \nu_n^*(\mathcal{H}) - 1\right) + O\left(\frac{1}{n}\right).$$

□

References

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. US National Bureau of Standards.

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

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

[4] Ash, R. B. and Doleans-Dade, C. (2000). *Probability and Measure Theory*. Academic Press.

[5] Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2016). PAC-Bayesian bounds based on the Rényi divergence. In *Proceedings of Machine Learning Research*, volume 51, pages 435–444.

[6] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

[7] Catoni, O. (2004). *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer.

[8] Catoni, O. (2007). *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. IMS Lecture Notes–Monograph Series. Institute of Mathematical Statistics.

[9] Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185.

[10] Catoni, O. and Giulini, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*.

[11] Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-gaussian mean estimators. *Annals of Statistics*, 44(6):2695–2725.

[12] Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016). PAC-Bayesian theory meets Bayesian Inference. In *Advances in Neural Information Processing Systems 29*, pages 1884–1892.

[13] Holland, M. J. (2019). Robust descent using smoothed multiplicative noise. In *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 703–711.

[14] Maurer, A. (2004). A note on the PAC Bayesian theorem. *arXiv preprint arXiv:cs/0411099*.

[15] McAllester, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363.

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

[17] McAllester, D. A. (2013). A PAC-Bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*.

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

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

[20] Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. (1996). A framework for structural risk minimisation. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 68–76. ACM.

[21] Tolstikhin, I. and Seldin, Y. (2013). PAC-Bayes-Empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems 26*, pages 109–117.

[22] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.