Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty

Felix Biggs
University College London and Inria London
contact@felixbiggs.com

Benjamin Guedj
University College London and Inria London
b.guedj@ucl.ac.uk

Abstract

We introduce a modified version of the excess risk, which can be used to obtain empirically tighter, faster-rate PAC-Bayesian generalisation bounds. This modified excess risk leverages information about the relative hardness of data examples to reduce the variance of its empirical counterpart, tightening the bound. We combine this with a new bound for $[-1, 1]$-valued (and potentially non-independent) signed losses, which is more favourable when they empirically have low variance around 0. The primary new technical tool is a novel result for sequences of interdependent random vectors which may be of independent interest. We empirically evaluate these new bounds on a number of real-world datasets.

1 INTRODUCTION AND OVERVIEW OF CONTRIBUTIONS

Generalisation bounds are of paramount importance in machine learning, both for understanding generalisation, and for obtaining guarantees for predictors. Obtaining the tightest possible bounds shines light on the former and leads to numerically better guarantees for the latter.

Consider a parameterised learning problem where we are interested in training a predictor h_w depending on weights w (*e.g.*, a neural network). In PAC-Bayes, predictions are typically made by drawing randomised weights $W \sim \rho$ where ρ is a so-called posterior distribution, then predicting $h_W(x)$ for some input x. Thus the learning is moved from the parameter w to a distribution ρ over W.

PAC-Bayesian generalisation bounds (Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999; Catoni, 2007) allow for quantifying the generalisation performance of predictors of the form h_W with high probability. They can also be used as a stepping stone to proving bounds where w is not random, for example for majority votes (Masegosa et al., 2020; Zantedeschi et al., 2021; Biggs et al., 2022). The recent surge in attention given to the PAC-Bayesian approach partially derives from a number of works establishing numerically non-vacuous bounds for neural networks with randomised (Dziugaite and Roy, 2017, 2018; Zhou et al., 2019; Letarte et al., 2019; Biggs and Guedj, 2021; Dziugaite et al., 2021; Perez-Ortiz et al., 2021b) or non-randomised (Biggs and Guedj, 2022) weights on real-world datasets. We refer to Guedj (2019) and Alquier (2021) and the many references therein for a broad introduction to PAC-Bayes.

Two terms commonly appear in PAC-Bayes bounds: $\mathrm{KL} := \mathrm{KL}(\rho, \pi)$, which defines the complexity of ρ as a Kullback-Leibler divergence from some sample-independent reference measure (usually referred to as "prior") π; and LG, a term logarithmic in the probability δ. If the number of examples is m, then at worst $\mathrm{LG} \leq \mathcal{O}(\log(m/\delta))$. The simplest such bound for bounded losses (McAllester, 1998) takes the form

$$\text{generalisation gap of } \rho \leq \mathcal{O}\left(\sqrt{\frac{\mathrm{KL} + \mathrm{LG}}{m}}\right),$$

holding with probability at least $1-\delta$ over the sample. The above is rarely tight, and was greatly improved by the bound of Maurer (2004), which we discuss further in Section 1.3. Maurer's bound has the advantage that it can (when the empirical loss of ρ is small) achieve a faster rate of convergence, where the dependence $\mathcal{O}(\sqrt{\mathrm{KL}/m})$ is improved to the "fast-rate" $\mathcal{O}(\mathrm{KL}/m)$. Since commonly $\mathrm{KL} \gg \mathrm{LG}$, this can lead to numerically tighter bounds.

A major question in (PAC-Bayesian) learning theory is *under what conditions such rates can be possible*.

As in VC theory, such fast-rates are possible when the empirical risk of ρ is zero, but it is also possible to get close to this fast regime under more general conditions. Getting such faster rates is a primary motivation for "Bernstein" and "Bennett"-type bounds (which leverage low variance

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

to get faster rates) in classical learning theory, as well as for the introduction of the excess loss, which combines nicely with the former.

1.1 NOTATION

In order to further discuss existing approaches, we define our terms more thoroughly. In the following, we examine different PAC-Bayesian generalisation bounds for bounded losses $\ell : \mathcal{W} \times \mathcal{Z} \to [0, 1]$ (where the specific range $[0, 1]$ is w.l.o.g. due to the possibility of rescaling). We let $\mathcal{W}$ denote the weight space and $\mathcal{Z}$ is the sample space.

A generalisation bound is an upper bound on the risk $\mathcal{L}(w) := \mathbb{E}_{Z\sim\mathcal{D}}\ell(w, Z)$, that holds for some data-dependent hypothesis w. We extend this by abuse of notation in a PAC-Bayesian setting to also write $\mathcal{L}(\rho) := \mathbb{E}_{W\sim\rho}\mathcal{L}(W)$ for PAC-Bayesian posterior distribution $\rho \in \mathcal{M}_1^+(\mathcal{W})$ (where $\mathcal{M}_1^+(\mathcal{A})$ denotes the space of probability measures on set $\mathcal{A}$).

We also introduce notation for a sequence of examples, $z_{1:i} = (z_1, \ldots, z_i) \in \mathcal{Z}^\star$, where $A^\star := \emptyset \cup \bigcup_{i=1}^\infty A^i$ is the set of sequences of elements in set A and we notate $z_{1:0} = \emptyset$. Learning takes place based on a i.i.d. sample of size m, $S = Z_{1:m} \sim \mathcal{D}^m$, and we define the empirical (in-sample) risk using it as $\widehat{\mathcal{L}}(w) = \mathbb{E}_{Z'\sim\mathrm{Uniform}(S)}\ell(w, Z')$.

1.2 FAST RATES AND EXCESS LOSSES

The simplest PAC-Bayesian bound which can achieve fast rates (and therefore tighter bound values) is the following:

$$\mathcal{L} - \widehat{\mathcal{L}} \leq \sqrt{\frac{\mathrm{KL} + \mathrm{LG}}{m} \cdot 2\widehat{\mathcal{L}}} + 2\frac{\mathrm{KL} + \mathrm{LG}}{m}. \tag{1}$$

This bound[1] (which is a relaxation of Maurer's bound, see Section 1.3) has a well-studied form common in classical learning theory where the KL term is replaced by a different complexity term. When $\widehat{\mathcal{L}} \to 0$ it achieves the fast rate on $\mathcal{L} - \widehat{\mathcal{L}}$ of $\mathcal{O}(\mathrm{KL}/m)$ and will be numerically tighter, but otherwise (for example, on a difficult dataset where $\widehat{\mathcal{L}}$ is large) the square root term typically dominates.

A common question in learning theory has therefore been on whether empirical risk under the square root can be replaced by something faster-decaying, like a variance (Tolstikhin and Seldin, 2013) or an *excess* risk. The excess risk (which we later generalise from this definition) is introduced by comparing the loss of our hypothesis w to a fixed "good" hypothesis $w^\star$ (we leave aside for now the question of choosing $w^\star$) in a modified loss function, $\tilde{\ell}(w, z) = \ell(w, z) - \ell(w^\star, z)$. This has the population and sample counterparts

$$\mathcal{E}(w) := \mathbb{E}_{Z\sim\mathcal{D}}\tilde{\ell}(w, Z) = \mathcal{L}(w) - \mathcal{L}(w^\star)$$

and

$$\widehat{\mathcal{E}}(w) := \mathbb{E}_{Z\sim\mathrm{Uniform}(S)}\tilde{\ell}(w, Z).$$

For example, Mhammedi et al. (2019) prove the "Unexpected Bernstein" PAC-Bayes bound

$$\mathcal{E} \leq \widehat{\mathcal{E}} + \mathcal{O}\left(\sqrt{\frac{\mathrm{KL} + \mathrm{LG}}{m} \cdot \widehat{V}} + \frac{\mathrm{KL} + \mathrm{LG}}{m}\right),$$

where $\widehat{V}(\rho) = \mathbb{E}_{W\sim\rho}[\frac{1}{m}\sum_{i=1}^m |\ell(W, Z_i) - \ell(w^\star, Z_i)|^2] \leq \widehat{\mathcal{E}}(\rho)$. The idea is that the second loss term in the excess risk "de-biases" and reduces the variance, so that if the predictors err on a similar set of examples, $\widehat{\mathcal{E}}(w)$ will be small, giving a faster rate. Such bounds on $\mathcal{E}$ can be converted back into generalisation bounds, by using that $\mathcal{L}(w^\star) - \widehat{\mathcal{L}}(w^\star) \leq \mathcal{O}(\sqrt{\mathrm{LG}/m})$ (since $w^\star$ is independent of the dataset) to get a bound like

$$\mathcal{L} \leq \widehat{\mathcal{L}} + \mathcal{O}\left(\sqrt{\frac{\mathrm{KL} + \mathrm{LG}}{m} \cdot \widehat{V}} + \frac{\mathrm{KL} + \mathrm{LG}}{m} + \sqrt{\frac{\mathrm{LG}}{m}}\right). \tag{2}$$

Since in most cases $\mathrm{LG} \ll \mathrm{KL}$, the final term is usually an insignificant price compared with the reduction from $\widehat{\mathcal{L}}$ to $\widehat{V}$. The rate of the final term can also be improved even further using assumptions about the noise (as examined at length in Mhammedi et al., 2019), or using dataset evaluations of the loss of $w^\star$.

A problem with this approach is the fact that $w^\star$ must be independent of the data. This means we must split the dataset as with PAC-Bayes data-dependent priors (as in, *e.g.*, Parrado-Hernández et al., 2012; Rivasplata et al., 2018; Mhammedi et al., 2019; Perez-Ortiz et al., 2021a), into parts used to produce $w^\star$ (and potentially learn a prior), and to actually apply the bound to. This reduces the effective sample size in the bound (*e.g.*, from m to $m/2$ when a 50-50 split is used). This issue can be partially circumvented through the use of forwards-backwards "informed" priors, but in expectation over different splits of the data this approach is actually weaker than the naive splitting procedure.

1.3 KL-BASED BOUNDS

The most well known (and often tightest) PAC-Bayesian bound for bounded losses $\in [0, 1]$ is Maurer's bound (Maurer, 2004):

$$\mathrm{kl}\left(\widehat{\mathcal{L}}(\rho) \| \mathcal{L}(\rho)\right) \leq \frac{\mathrm{KL}(\rho, \pi) + \log\frac{2\sqrt{m}}{\delta}}{m}$$

where $\mathrm{kl}(q\|p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$ is the KL divergence between Bernoulli distributions of biases q, p. This

[1] Note that for the sake of clarity we will make the slight notational abuse of omitting the argument of $\mathcal{L}$, $\widehat{\mathcal{L}}$, $\mathcal{E}$ and $\widehat{\mathcal{E}}$ when the context is clear.

bound can be inverted to obtain an upper bound directly on $\mathcal{L}$ by defining the inverse

$$\mathrm{kl}^{-1}(u\|b) := \sup\{r : \mathrm{kl}(u\|r) \leq b\}.$$

The bound in Equation (1) is obtained through the relaxation $\mathrm{kl}^{-1}(u\|b) \leq u + \sqrt{2bu} + 2b$ (McAllester, 2003). However, note that this lower bound can be considerably weaker, as it does not leverage the combinatorial power of the small-kl. It has been shown (Foong et al., 2021) that no bound on the naive loss (but not necessarily when we leverage the excess loss) can improve Maurer's bound (aside from the open question of whether it is possible to remove the logarithmic factor).

We note also that although the small-kl bound can be rescaled to use the excess loss, this leads to a bound like Equation (2) with $\widehat{V} = \frac{\widehat{\mathcal{E}}+1}{2} \geq \frac{1}{2}\widehat{\mathcal{L}}$ (by applying the bound to $\frac{\mathcal{E}+1}{2}$ and relaxing as above). This does not lead to fast rates under different conditions to usual, since it is still necessary that $\widehat{\mathcal{L}} \to 0$.

Recently, Adams et al. (2022) proved a generalisation of this bound which holds for *vector*-valued losses, $\boldsymbol{\ell} : \mathcal{W} \times \mathcal{Z} \to \Delta_M$ (with Δ_M the M-dimensional simplex),

$$\mathrm{kl}\left(\widehat{\boldsymbol{U}}\|\boldsymbol{\mu}\right) \leq \frac{\mathrm{KL}(\rho,\pi) + \log\frac{\xi(M,m)}{\delta}}{m}.$$

where $\widehat{U}_i(w) = \frac{1}{m}\sum_{i=1}^m \ell_i(w,z)$ and $\boldsymbol{\mu} = \mathbb{E}_S\widehat{\boldsymbol{U}}$, and $\xi(M,m)$ is a polynomial function of m. Inverting such a bound is somewhat more complicated, but we can use it to obtain an upper bound on $\sum_i \alpha_i U_i$ for some set of coefficients α_i. This is the tool we will use to obtain our bounds.

1.4 OUR CONTRIBUTIONS

From the above starting points, we pursue two parallel and complimentary directions of improvement. First we provide a generalisation of the excess risk which allows $w^\star$ to be learned from the stream of data as we receive it. In this way, we are able to more effectively "de-bias" our bounds, reducing the effective variance term. This is intended to demonstrate how PAC-Bayes bounds can be made tighter by using information from the training set (in a sense, the relative difficulty of different examples) more efficiently.

Secondly, we observe that Equation (1) is a *relaxation* of Maurer's bound, and this weakening leads to a loss of some of the tightness of the original "kl" formulation. Thus we give a new bound which leverages the tightness of kl-based bounds like Maurer's, but also relaxes to a form like that in Equation (2). Specifically, it reduces to the form given in Equation (2) with $\widehat{V}(\rho) = \mathbb{E}_{W\sim\rho}\frac{1}{m}|\ell(W,Z_i) - \ell(w^\star,Z_i)|$. This is very similar (if slightly larger) to the term given in Mhammedi et al. (2019), although both are equivalent when $\ell \in \{0,1\}$, as for example with the misclassification loss, which ensures faster rates under similar conditions.

However this form of our bound is only a relaxation, and the kl-type formulation that we give for it is considerably tighter empirically, and we show it is strictly tighter under a $\{0,1\}$-valued loss. In combination, these lead to a new bound on the out-of-sample risk that is empirically tighter than Maurer's bound in some cases, a feat very rare in the literature.

In order to prove these results, we make several technical contributions, including new results for simplex-valued Martingales within PAC-Bayes and a new method for relaxing our kl-type bound. We also note that our bound can improve that of Wu and Seldin (2022) for ternary $\in \{-1,0,1\}$ random variables.

2 WARM UP

In this section we give a simplified version of our main results, discussing only classification. In this setting, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ for $\mathcal{Y} = \{1,\ldots,c\}$, and $S = \{(X_i,Y_i)\}_{i=1}^m$. Our predictions are given by $h_w(x)$ for $w \in \mathcal{W}$ and we consider the misclassification loss, $\ell(w,(x,y)) = \mathbf{1}_{h_w(x)\neq y}$.

We consider positive and negative deviations of the excess loss, $\widehat{\mathcal{E}}_+$ and $\widehat{\mathcal{E}}_-$, which we will define for general losses in the next section. For intuition, in the *specific case* of misclassification loss they are equal to the following

$$\widehat{\mathcal{E}}_+^{\mathrm{mc}} = \mathbb{E}_{W\sim\rho}\left[\frac{|\{(x,y)\in S : h_W(x)\neq y, h_{w^\star}(x) = y\}|}{m}\right],$$
$$\widehat{\mathcal{E}}_-^{\mathrm{mc}} = \mathbb{E}_{W\sim\rho}\left[\frac{|\{(x,y)\in S : h_W(x)= y, h_{w^\star}(x) \neq y\}|}{m}\right].$$

Thus in this case we are merely counting the numbers of two different types of loss: an error using w but not using $w^\star$, and the converse. If neither predictor or both predictors err, this incurs no loss. These two error types have simple interpretations as counts, which is similar to the work of Adams et al. (2022), and indeed the preliminary results we show here could follow fairly straightforwardly from theirs.

We note that $\mathcal{E} = \mathcal{E}_+ - \mathcal{E}_-$, and

$$\widehat{\mathcal{E}}_+^{\mathrm{mc}} + \widehat{\mathcal{E}}_-^{\mathrm{mc}} = \frac{1}{m}\sum_{i=1}^m \mathbb{P}_{W\sim\rho}(h_W(X_i) \neq h_{w^\star}(X_i)),$$

a form which commonly appears in learning theory, and can also be controlled by Mammen-Tsybakov or Massart noise conditions.

Our main result implies the following (weakened) relaxed bound:

$$\mathcal{E}(\rho) \leq \widehat{\mathcal{E}}(\rho) + 2\sqrt{\frac{\mathrm{KL}+\mathrm{LG}}{m}\cdot(\widehat{\mathcal{E}}_+ + \widehat{\mathcal{E}}_-)} + 2\frac{\mathrm{KL}+\mathrm{LG}}{m} \tag{3}$$

with $\mathrm{LG} = \log(2m/\delta)$.

We then can easily go from Equation (3) to a bound on the population risk (like Equation (2)), since $w^\star$ is independent of the data, and can be estimated empirically with no complexity penalty. The (simplest) Hoeffding bound gives $\mathcal{L}(w^\star) - \widehat{\mathcal{L}}(w^\star) \leq \sqrt{\log(1/\delta)/2m}$. Overall, with cancellations (and a union bound, so that the following holds with probability $1-2\delta$) this gives a bound on $\mathcal{L}(\rho) - \widehat{\mathcal{L}}(\rho)$ of

$$2\sqrt{\frac{\mathrm{KL}+\mathrm{LG}}{m}\cdot(\widehat{\mathcal{E}}_+ + \widehat{\mathcal{E}}_-)} + 2\frac{\mathrm{KL}+\mathrm{LG}}{m} + \sqrt{\frac{\log \delta^{-1}}{2m}}. \quad (4)$$

The last term is complexity-free and hence will generally be dominated by the other terms.

2.1 KL-TYPE FORMULATION

Here we show that the above can be considerably tightened into a "kl"-type formulation. This can be relaxed back into the previous form, similarly to the way in which Maurer's bound implies the much weaker result in Equation (1). Collecting the error type counts into the vector

$$\widehat{\boldsymbol{\mathcal{E}}}(\rho) = [\widehat{\mathcal{E}}_+(\rho), \widehat{\mathcal{E}}_-(\rho), 1 - \widehat{\mathcal{E}}_+(\rho) - \widehat{\mathcal{E}}_-(\rho)]^T,$$

we will prove the bound

$$\mathcal{E}(\rho) \leq \phi\left(\widehat{\boldsymbol{\mathcal{E}}}(\rho), \frac{\mathrm{KL}(\rho,\pi) + \log\frac{2m}{\delta}}{m}\right)$$

where $\phi(\boldsymbol{u}, b) := \sup\{r_1 - r_2 : \boldsymbol{r} \in \Delta_3, \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{r}) \leq b\}$.

This function (and its gradients) can be calculated by a simple procedure outlined in Appendix B.2 This form leverages the tightness of the kl bound and is empirically much tighter than Equation (3). In the next section we give a more detailed analysis of this function and show that it implies several relaxations, including Equation (3), the unexpected Bernstein bound (for certain losses), and a split-kl type inequality similar to that from Wu and Seldin (2022).

For intuition, we point out that this basic result (though not the relaxations which require further proofs, or the general bounded loss forms, since their result applies only for count-based losses) can be proved by application of results from Adams et al. (2022) to a vector loss of counts, $\boldsymbol{\ell} : \mathcal{W} \times \mathcal{Z} \to \Delta_3$, with

$$\begin{aligned}\ell_1 &= \mathbf{1}_{h_w \neq y, h_{w^\star} = y},\\ \ell_2 &= \mathbf{1}_{h_w = y, h_{w^\star} \neq y},\\ \ell_3 &= 1 - \ell_1 - \ell_2.\end{aligned}$$

Just as above, this result can be combined with a (test set) bound on $\mathcal{L}(w^\star)$ using $\widehat{\mathcal{L}}(w^\star)$ to provide a generalisation bound for ρ, as in Equations (2) and (4). Again, to leverage the tightness of the kl formulation, we could replace the weaker Hoeffding bound used before with a tighter (kl-based, inverted) Chernoff bound (as given in e.g. Foong et al., 2022):

$$\mathcal{L}(w^\star) \leq \mathrm{kl}^{-1}\left(\widehat{\mathcal{L}}(w^\star) \middle\| \frac{\log\frac{1}{\delta}}{m}\right).$$

Thus we can get an overall result with two inverse kl-type terms, which is much tighter in practice than the formulations given in the previous section. In Section 3 we give formal statements of the above and adapt the bound to work for any bounded loss function.

2.2 GENERALISING THE EXCESS LOSS

A problem with formulations using the excess risk is that the optimal choice of $w^\star$ is essentially impossible to know in advance. In order to optimally de-bias bounds, $w^\star$ ought to perform similarly to our overall posterior, essentially identifying difficult examples, but this is very difficult when it must be chosen in a data-free way. One solution to this problem is data-splitting, but this is sub-optimal because it reduces the amount of data available to the bound.

Here instead, drawing on PAC-Bayesian tools for non-independent, martingale based losses, we propose an alternative solution through a generalisation of the excess risk. We first recall the notation $z_{1:i}$ for a sequence of examples $z_1, \ldots, z_i$. Given a sequence of i.i.d. variables Z_i, it turns out we can derive empirically-calculable bounds by considering a sequence of losses like

$$\ell(w, Z_i) - \ell(w_i^\star(Z_{1:i-1}), Z_i). \quad (5)$$

Here, instead of being fixed in a data independent way, $w_i^\star(Z_{1:i-1})$ is an online sequence of weights learned using the first $i-1$ examples. If the procedure used to learn the $w_i^\star$ is similar to that used to learn $W \sim \rho$, and is relatively stable to changes in dataset size, the errors of ρ and the $w_i^\star$ will be highly correlated. The terms in Equation (5) will mostly cancel, reducing the excess risk and tightening the overall bound. This approach can be easily generalised to stochastic online algorithms, a procedure mentioned by Mhammedi et al. (2019) as "online estimators", but not used empirically. We note that, unlike with data-dependent priors in PAC-Bayes bounds, no data-splitting is necessary for this procedure, the bound still uses all of the training data with $m = |S|$.

To clarify further, this technique can tighten any bound that leverages excess losses to achieve faster rates, such as ours or the unexpected Bernstein, but *not* Maurer's bound. Ordering examples by difficulty is not necessary; what is useful is that a learner on the first i points makes similar errors to our (whole-dataset) posterior, leveraging the notion that not all examples have the same difficulty. This is what we mean by "leveraging example difficulty".

3 MAIN RESULTS

In the following, $\ell : \mathcal{W} \times \mathcal{Z} \to [0,1]$ is a bounded loss and $S = Z_{1:m}$ is an i.i.d. sample from unknown distribution $\mathcal{D}$. We choose a sequence of online estimators $\rho_i^\star \in \mathcal{M}_1^+(\mathcal{W})$ for $i \in \{1, \dots, m\}$, each depending on only the previous $i-1$ examples $Z_{1:i-1}$. This might be done for example through choosing $\rho_i^\star = \mathcal{A}(Z_{1:i-1})$, where $\mathcal{A} : \mathcal{Z}^\star \to \mathcal{M}_1^+(\mathcal{W})$ is a predetermined algorithm.

Our de-biased (generalised, in that it depends on i examples) loss for the ith example takes the form

$$\tilde{\ell}(w, z_{1:i}) := \ell(w, z_i) - \mathbb{E}_{W' \sim \rho_i^\star}[\ell(W', z_i)].$$

We distinguish two different types of sample errors (corresponding to positive and negative parts $\tilde{\ell} > 0$ and $\tilde{\ell} < 0$) with corresponding generalised empirical risk values:

$$\widehat{\mathcal{E}}_+(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m \max(\tilde{\ell}(W, Z_{1:i}), 0)\right],$$
$$\widehat{\mathcal{E}}_-(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m \max(-\tilde{\ell}(W, Z_{1:i}), 0)\right].$$

For notational convenience we collect these in the vector

$$\widehat{\boldsymbol{\mathcal{E}}}(\rho) = [\widehat{\mathcal{E}}_+(\rho), \widehat{\mathcal{E}}_-(\rho), 1 - \widehat{\mathcal{E}}_+(\rho) - \widehat{\mathcal{E}}_-(\rho)]^T.$$

Given a PAC-Bayesian posterior $\rho \in \mathcal{M}_1^+(\mathcal{W})$, we define the generalised excess risk as

$$\mathcal{E}(\rho) := \mathcal{L}(\rho) - \mathcal{L}^\star$$

with

$$\mathcal{L}^\star := \frac{1}{m}\sum_{i=1}^m \mathbb{E}_Z \mathbb{E}_{W\sim\rho_i^\star}[\ell(W, Z)].$$

We next present a result in two parts which can be used to control $\mathcal{E}$ and $\mathcal{L}^\star$, and hence obtain a bound on $\mathcal{L}(\rho)$ directly.

Theorem 1 (Generalisation Loss Bound)**.** *Fix data distribution $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{Z})$, prior $\pi \in \mathcal{M}_1^+(\pi)$, $m \geq 3, \delta \in (0,1)$, and bounded loss $\ell : \mathcal{Z} \times \mathcal{W} \to [0,1]$. Let $S = Z_{1:m} \sim \mathcal{D}^m$ be a sample and $\rho_i^\star \in \mathcal{M}_1^+(\mathcal{W})$ be any sequence of online estimators with $i = 1, \dots, m$, each depending only on the $i-1$ examples $Z_{1:i-1}$, and let $\mathcal{E}$ and $\widehat{\boldsymbol{\mathcal{E}}}$ be as defined in this section.*

Then, with probability at least $1-\delta$ over S for arbitrary posterior $\rho \in \mathcal{M}_1^+(\mathcal{W})$,

$$\mathcal{E}(\rho) \leq \phi\left(\widehat{\boldsymbol{\mathcal{E}}}(\rho), \frac{\mathrm{KL}(\rho,\pi) + \log\frac{2m}{\delta}}{m}\right),$$

where $\phi(\boldsymbol{u}, b) := \sup\{r_1 - r_2 : \boldsymbol{r} \in \Delta_3, \mathrm{kl}(\boldsymbol{u}\|\boldsymbol{r}) \leq b\}$.

Further, with probability at least $1-\delta$ over S,

$$\mathcal{L}^\star \leq \mathrm{kl}^{-1}\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho_i^\star}[\ell(W, Z_i)] \middle\| \frac{\log\frac{1}{\delta}}{m}\right).$$

With probability at least $1-2\delta$, the above hold simultaneously, and $\mathcal{L}(\rho) = \mathcal{E}(\rho) + \mathcal{L}^\star$ is bounded by the sum of the right hand sides.

This is a complex result, so to begin we note that a deterministic and data-free choice of $\rho_i^\star$ as a distribution fixed to $w^\star$ recovers the results from Section 2.1. The relaxed form given by Equation (3) is still also valid with the modified forms of the excess risk, and gives intuition about the bound; in this more general (bounded, not misclassification) case,

$$\widehat{\mathcal{E}}_+ + \widehat{\mathcal{E}}_- = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho}|\ell(W, Z_i) - \mathbb{E}_{W'\sim\rho_i^\star}\ell(W', Z_i)|.$$

In the next sub-sections, we examine some corollaries and relaxations of this bound. We also note here that it is possible to directly calculate the value and gradients of ϕ with respect to both of its arguments using a procedure outlined by Adams et al. (2022), which could be very useful in optimising the bound directly as an objective.

3.1 RELAXATION TO MAURER

As a basic sanity check to show that the new bound leverages the tightness of a small-kl-based bound, we show that it can be used to recover Maurer's bound (up to a factor 2 in front of the logarithmic term). Assume the realisable case[2], where there exists $w^\star$ such that $\ell(w^\star, z) = 0$ for all z in the support of $\mathcal{D}$. This ensures that $\widehat{\mathcal{E}}_- = 0$, $\mathcal{L}(\rho) = \mathcal{E}(\rho)$, and

$$\widehat{\mathcal{E}}_+(\rho) = \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m \ell(W, Z_i)\right] = \widehat{\mathcal{L}}(\rho).$$

We also have the following.

Proposition 1. *For any $u \in [0,1]$ and $b > 0$*

$$\phi([u, 0, 1-u], b) = \mathrm{kl}^{-1}(u\|b).$$

Combining Proposition 1 with our Theorem 1 in this setting implies that

$$\mathcal{L}(\rho) \leq \mathrm{kl}^{-1}\left(\widehat{\mathcal{L}}(\rho) \middle\| \frac{\mathrm{KL}(\rho,\pi) + \log\frac{2m}{\delta}}{m}\right),$$

which is Maurer's bound (up to a constant factor for the log term). In the realisable case, we would expect $\widehat{\mathcal{L}}(\rho) \to 0$

[2] Surprisingly, we can assume this w.l.o.g. by a mathematical trick, extending $\mathcal{W}$ to contain a special, non-learnable point $w^\dagger$ such that $\ell(w^\dagger, z) = 0$ for all z. However, this deprives us of the use of excess losses, which can give tighter bounds when we do not learn a low-risk solution as in the "true" realisable case.

and de-biasing is not necessary to achieve faster rates, so this shows that we can recover the tightest known bound for that scenario.

3.2 RELAXATION TO UNEXPECTED BERNSTEIN AND NOISE CONDITIONS

The following result can be used to obtain a number of relaxations of our result that enable more direct comparison to Mhammedi et al. (2019).

Proposition 2. *For any $\boldsymbol{u} \in \Delta_3, b > 0$,*

$$\begin{aligned}\phi(\boldsymbol{u}, b) &\leq \inf_{\eta \in (0,1)} \frac{1}{\eta}\left(1 - e^{u_1 \log(1-\eta) + u_2 \log(1+\eta) - b}\right) \\ &\leq \inf_{\eta \in (0,1)} f_\eta(u_1 - u_2 + c_\eta(u_1 + u_2) + b/\eta)\end{aligned}$$

where $f_\eta(t) = \eta^{-1}(1 - e^{-\eta t}) \leq t$, and $c_\eta = -1 - \log(1-\eta)/\eta$ is a term also appearing in the unexpected Bernstein bound.

This enables us to relax our main result to find that simultaneously for any $\eta \in (0, 1)$

$$\mathcal{E} \leq f_\eta\left(\widehat{\mathcal{E}}_+ - \widehat{\mathcal{E}}_- + c_\eta(\widehat{\mathcal{E}}_+ + \widehat{\mathcal{E}}_-) + \frac{\mathrm{KL} + \mathrm{LG}}{m\eta}\right). \quad (6)$$

Firstly we note that setting $\eta = 1 - e^{-C}$ for $C > 0$ gives a bound similar to Catoni's Catoni (2007) with a free choice of C and an extra $\log(2m)$ term, and recovers it exactly when $\widehat{\mathcal{E}}_- = 0$.

We also compare this relaxation with the strongest form of the unexpected Bernstein (given by Mhammedi et al. (2019) and Theorem 6 our appendix), which takes the form

$$\mathcal{E} \leq \widehat{\mathcal{E}}_+ - \widehat{\mathcal{E}}_- + c_\eta \hat{V} + \frac{\mathrm{KL} + \mathrm{LG}}{m\eta}$$

for a relatively free (it must be chosen from within some covering grid) choice of η. In the specific case of the misclassification loss (and non-randomised $\rho_i^\star$ each supported on a single point), the term $\widehat{\mathcal{E}}_+ + \widehat{\mathcal{E}}_- = \hat{V}$, is equal to that from the unexpected Bernstein. We therefore recover a bound (Equation (6)) taking the same form but with an extra $f_\eta(t) \leq t$ around it, so in the case of misclassification loss our bound is always at least as strong as the results of Mhammedi et al. (2019). For more general losses, since the squaring of the loss difference terms in $\hat{V}$ makes them smaller, the unexpected Bernstein might be more able to leverage small but non-zero per-example excess losses.

A corollary of this misclassification equivalence is that our bound can achieve faster rates under a (c, β)-Mammen-Tsybakov noise condition. These noise conditions for the misclassification loss imply a β-Bernstein condition. Under the analysis of Mhammedi et al. (2019, Section 5), with the same learning algorithm choices, our bound therefore achieves a rate of at least $\tilde{O}(m^{-1/(2-\beta)})$.

3.3 RELAXATION TO SPLIT-KL

A relaxation of our bound that is very similar in form to the recent PAC-Bayes split-kl inequality from Wu and Seldin (2022) is obtained through the following proposition.

Proposition 3. *For any $\boldsymbol{u} \in \Delta_3$, $b > 0$,*

$$\begin{aligned}\phi(\boldsymbol{u}, b) &\leq \mathrm{kl}^{-1}(u_1 \| b) - \mathrm{kl}_{\mathrm{LB}}^{-1}(u_2 \| b) \\ &\leq u_1 - u_2 + 2\sqrt{b \cdot (u_1 + u_2)} + 2b,\end{aligned}$$

where $\mathrm{kl}_{\mathrm{LB}}^{-1}(u \| b) := \inf\{r \in [0,1] : \mathrm{kl}(u \| r) \leq b\}$ is the lower *tail small-kl inversion.*

This gives the bound

$$\mathcal{E} \leq \mathrm{kl}^{-1}\left(\widehat{\mathcal{E}}_+ \middle\| \frac{\mathrm{KL} + \mathrm{LG}}{m}\right) - \mathrm{kl}_{\mathrm{LB}}^{-1}\left(\widehat{\mathcal{E}}_- \middle\| \frac{\mathrm{KL} + \mathrm{LG}}{m}\right). \quad (7)$$

This is essentially the same as the split-kl inequality (Wu and Seldin, 2022), except that we have $\mathrm{LG} = \log \frac{2m}{\delta}$ while in their bound $\mathrm{LG} = \log \frac{4\sqrt{m}}{\delta}$, so the constants in ours are slightly worse, as in Section 3.1. Their main bound is not limited to excess losses, but is primarily aimed at losses where there are three different special values to be focused on (in the simple case, $\{-1, 0, 1\}$, as here). We note that the techniques they use to do this (which involve re-scaling and translating loss functions) could also be combined with our (non-split) kl formulation.

3.4 ASIDE: SLIGHT GENERALISATIONS

We here give a slight generalisation of our main result. This suggests two different possible directions for our main technical results.

Theorem 2. *For any measurable $\tilde{\ell} : \mathcal{W} \times \mathcal{Z}^\star \to [-1, 1]$ as defined above and any $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{Z}), m \geq 3$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$ simultaneously for all $\rho \in \mathcal{M}_1^+(\mathcal{W})$,*

$$\mathfrak{L}(\rho) \leq \phi\left(\widehat{\mathfrak{L}}(\rho), \frac{\mathrm{KL}(\rho, \pi) + \log \frac{2m}{\delta}}{m}\right),$$

where we have defined

$$\begin{aligned}\widehat{\mathfrak{L}}_+(\rho) &:= \mathbb{E}_{W \sim \rho}\left[\frac{1}{m}\sum_{i=1}^m \max(\tilde{\ell}(W, Z_{1:i}), 0)\right], \\ \widehat{\mathfrak{L}}_-(\rho) &:= \mathbb{E}_{W \sim \rho}\left[\frac{1}{m}\sum_{i=1}^m \max(-\tilde{\ell}(W, Z_{1:i}), 0)\right], \\ \widehat{\mathfrak{L}}(\rho) &:= [\widehat{\mathfrak{L}}_+(\rho), \widehat{\mathfrak{L}}_-(\rho), 1 - \widehat{\mathfrak{L}}_+(\rho) - \widehat{\mathfrak{L}}_-(\rho)]^T, \\ \mathfrak{L}(\rho) &:= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{Z_i}[\tilde{\ell}(Z_{1:i}) | Z_{1:i-1}],\end{aligned}$$

and $\phi(\boldsymbol{u}, b) := \sup\{r_1 - r_2 : \boldsymbol{r} \in \Delta_3, \mathrm{kl}(\boldsymbol{u} \| \boldsymbol{r}) \leq b\}$.

Firstly, this shows that we can straightforwardly consider general "signed" loss in $[-1,1]$, rather than an excess loss. This takes our work closer to that of Wu and Seldin (2022).

We can also consider a more general de-biasing term, setting $\tilde{\ell} = \ell(w,z_i) - \ell^\star(z_{1:i})$, where $\ell^\star : \mathcal{Z}^\star \to [0,1]$. However, to obtain bounds on $\mathcal{L}(\rho)$ with for arbitrary choices of this function, we must also bound

$$\mathcal{L}(\rho) - \mathfrak{L}(\rho) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{Z_i}[\ell^\star(Z_{1:i})|Z_{1:i-1}].$$

This is difficult in general, but in some cases the following result is useful: One way to do this is suggested by the following:

Theorem 3 (Martingale Chernoff-Hoeffding Inversion)**.** *Let $U_i \in [0,1], i = 1,\dots,m$ have conditional expectations $\mathbb{E}[U_i|U_{1:i-1}] = \mu_i$, and averages $\bar{U} := \frac{1}{m}\sum_{i=1}^m U_i$, $\bar{\mu} = \frac{1}{m}\sum_{i=1}^m \mu_i$. Then with probability at least $1-\delta$,*

$$\bar{\mu} \leq \mathrm{kl}^{-1}\left(\bar{U}\middle\| \frac{\log\frac{1}{\delta}}{m}\right).$$

This result shows that any fixed method of choosing the de-biasing term can work. For example, we could train some model on the first $i-1$ points to predict the loss (which is in $[0,1]$) of the next data point. As long as this model is not trained on the i-th point, its evaluation on this point is akin to a real test-set evaluation, and so we can get a relatively sharp bound on the term $\mathcal{L}(\rho) - \mathfrak{L}(\rho)$.

4 PROOFS AND COROLLARIES

Firstly, we prove two theorems which generalise theorems of Adams et al. (2022) to random variables with a dependence structure, using ideas from Seldin et al. (2012). These results may be of interest in their own right.

Theorem 4 (Generalisation of Lemma 5 in Adams et al., 2022 and Lemma 1 in Seldin et al., 2012)**.** *Let $\boldsymbol{U}_1,\dots,\boldsymbol{U}_m$ be a sequence of random vectors, each in Δ_M, such that*

$$\mathbb{E}[\boldsymbol{U}_i|\boldsymbol{U}_1,\dots,\boldsymbol{U}_{i-1}] = \boldsymbol{\mu}_i$$

for $i = 1,\dots,m$. Let $\boldsymbol{V}_1,\dots,\boldsymbol{V}_m$ be independent Multinomial$(1,M,\boldsymbol{\mu}_i)$ *random vectors such that $\mathbb{E}\boldsymbol{V}_i = \boldsymbol{\mu}_i$. Then for any convex function $f : \Delta_M^m \to \mathbb{R}$:*

$$\mathbb{E}[f(\boldsymbol{U}_1,\dots,\boldsymbol{U}_m)] \leq \mathbb{E}[f(\boldsymbol{V}_1,\dots,\boldsymbol{V}_m)].$$

Proof. Let E_M denote the set of canonical (axis-aligned) M-dimensional basis vectors, for example $E_3 = \{[1,0,0],[0,1,0],[0,0,1]\}$. We will denote typical members of this set by $\boldsymbol{\eta}_i$, and tuples $\boldsymbol{\eta}_{1:m} = (\boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_m) \in \Delta_M^m$. Firstly we show that the definitions in the theorem lead to a Martingale-type result:

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=1}^m \boldsymbol{U}_i\cdot\boldsymbol{\eta}_i\right] \\
&= \mathbb{E}_{\boldsymbol{U}_{1:m-1}}\left[\left(\prod_{i=1}^{m-1}\boldsymbol{U}_i\cdot\boldsymbol{\eta}_i\right)\mathbb{E}_{\boldsymbol{U}_m}\left[\boldsymbol{U}_m|\boldsymbol{U}_{1:m-1}\right]\cdot\boldsymbol{\eta}_m\right] \\
&= \mathbb{E}_{\boldsymbol{U}_{1:m-1}}\left[\left(\prod_{i=1}^{m-1}\boldsymbol{U}_i\cdot\boldsymbol{\eta}_i\right)\boldsymbol{\mu}_m\cdot\boldsymbol{\eta}_m\right] \\
&= \prod_{i=1}^m \boldsymbol{\mu}_i\cdot\boldsymbol{\eta}_i.
\end{aligned}$$

In Adams et al. (2022, proof of Lemma 5), it is shown that for any convex function $f : \Delta_M^m \to \mathbb{R}$ and $\boldsymbol{u}_{1:m} = (\boldsymbol{u}_1,\dots,\boldsymbol{u}_m) \in \Delta_M^m$,

$$f(\boldsymbol{u}_{1:m}) \leq \sum_{\boldsymbol{\eta}_{1:m}\in E_M^m}\left(\prod_{i=1}^m \boldsymbol{u}_i\cdot\boldsymbol{\eta}_i\right) f(\boldsymbol{\eta}_{1:m}).$$

Applying this result to the random variables $\boldsymbol{U}_{1:m}$ and combining with the Martingale-type result leads to the following:

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{U}_{1:m})] \\
&\leq \mathbb{E}\left[\sum_{\boldsymbol{\eta}_{1:m}\in E_M^m}\left(\prod_{i=1}^m \boldsymbol{U}_i\cdot\boldsymbol{\eta}_i\right) f(\boldsymbol{\eta}_{1:m})\right] \\
&= \sum_{\boldsymbol{\eta}_{1:m}\in E_M^m}\mathbb{E}\left[\prod_{i=1}^m \boldsymbol{U}_i\cdot\boldsymbol{\eta}_i\right] f(\boldsymbol{\eta}_{1:m}) \\
&= \sum_{\boldsymbol{\eta}_{1:m}\in E_M^m}\left(\prod_{i=1}^m \boldsymbol{\mu}_i\cdot\boldsymbol{\eta}_i\right) f(\boldsymbol{\eta}_{1:m}) \\
&= \sum_{\boldsymbol{\eta}_{1:m}\in E_M^m}\left(\prod_{i=1}^m \mathbb{P}(\boldsymbol{V}_i = \boldsymbol{\eta}_i)\right) f(\boldsymbol{\eta}_{1:m}) \\
&= \mathbb{E}f(\boldsymbol{V}_{1:m}).
\end{aligned}$$

The final step in the proof followed via the definition of expectation w.r.t. the $\boldsymbol{V}_i$. □

Theorem 5 (Martingale PAC-Bayes for Vector KL.)**.** *Let $\boldsymbol{U}_1(w),\dots,\boldsymbol{U}_m(w)$ be a sequence of random vector valued functions, each in Δ_M, such that*

$$\mathbb{E}[\boldsymbol{U}_i(w)|\boldsymbol{U}_1(w),\dots,\boldsymbol{U}_{i-1}(w)] = \boldsymbol{\mu}_i(w)$$

for $i = 1,\dots,m$ and all $w \in \mathcal{W}$. Define

$$\widehat{\boldsymbol{U}}(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m \boldsymbol{U}_i(W)\right]$$

and

$$\bar{\boldsymbol{\mu}}(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m \boldsymbol{\mu}_i(W)\right].$$

Then for fixed $\pi \in \mathcal{M}_1^+(\mathcal{W}), \delta \in (0,1)$, with probability at least $1-\delta$ (over $\{\boldsymbol{U}_i(w) : i \in 1,\dots,m : w \in \mathcal{W}\}$), simultaneously for all $\rho \in \mathcal{M}_1^+(\mathcal{W})$,

$$\mathrm{kl}\left(\widehat{\boldsymbol{U}}(\rho)\middle\|\bar{\boldsymbol{\mu}}(\rho)\right) \leq \frac{\mathrm{KL}(\rho,\pi) + \log\frac{\xi(M,m)}{\delta}}{m}$$

where $\xi(M, m)$ is defined for $m \geq M$ by

$$\sqrt{\pi} e^{1/12m} \left(\frac{m}{2}\right)^{\frac{M-1}{2}} \sum_{k=0}^{M-1} \binom{M}{k} \frac{1}{(m\pi)^{k/2} \Gamma\left(\frac{M-k}{2}\right)}.$$

Proof. The proof begins by a common pattern in PAC-Bayesian proofs (as first pointed out by Germain et al., 2009). By Jensen's inequality, the Donsker-Varadhan change-of-measure theorem, Markov's inequality and the independence of π from $\{\boldsymbol{U}_i, \boldsymbol{\mu_i}\}$, the following holds with at least $1-\delta$ for any ρ:

$$\begin{aligned}
&m \operatorname{kl}\left(\widehat{\boldsymbol{U}}(\rho) \middle\| \bar{\boldsymbol{\mu}}(\rho)\right) - \operatorname{KL}(\rho, \pi) \\
&\leq m \mathbb{E}_{W\sim\rho} \operatorname{kl}\left(\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{U}_i(w) \middle\| \bar{\boldsymbol{\mu}}(w)\right) - \operatorname{KL}(\rho, \pi) \\
&\leq \log \mathbb{E}_{W\sim\pi}\left[e^{m \operatorname{kl}\left(\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{U}_i(w) \| \bar{\boldsymbol{\mu}}(w)\right)}\right] \\
&\leq \log \frac{1}{\delta} \mathbb{E}_{\boldsymbol{U}_i} \mathbb{E}_{W\sim\pi}\left[e^{m \operatorname{kl}\left(\frac{1}{m}\sum_i \boldsymbol{U}_i(w) \| \bar{\boldsymbol{\mu}}(w)\right)}\right] \\
&\leq \log \frac{1}{\delta} \mathbb{E}_{W\sim\pi} \mathbb{E}_{\boldsymbol{U}_i}\left[e^{m \operatorname{kl}\left(\frac{1}{m}\sum_i \boldsymbol{U}_i(w) \| \bar{\boldsymbol{\mu}}(w)\right)}\right].
\end{aligned}$$

By applying Theorem 4 to the inner term we find that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{U}_i}\left[e^{m \operatorname{kl}\left(\frac{1}{m}\sum_i \boldsymbol{U}_i(w) \| \bar{\boldsymbol{\mu}}(w)\right)}\right] \\
&\quad\leq \mathbb{E}_{\boldsymbol{V}_i}\left[e^{m \operatorname{kl}\left(\frac{1}{m}\sum_i \boldsymbol{V}_i(w) \| \bar{\boldsymbol{\mu}}(w)\right)}\right] \\
&\quad\leq \mathbb{E}_{\bar{\boldsymbol{V}}}\left[e^{m \operatorname{kl}\left(\bar{\boldsymbol{V}} \| \bar{\boldsymbol{\mu}}(w)\right)}\right],
\end{aligned}$$

where $\bar{\boldsymbol{V}} \sim \text{Multinomial}(m, M, \bar{\boldsymbol{\mu}}(w))$. The latter step follows as the expectation of a convex sum of Multinomial variables is maximised by variables having the same constants, $\boldsymbol{\mu}_i = \bar{\boldsymbol{\mu}}$ (Hoeffding, 1956). This final term is shown in Corollary 7 of Adams et al. (2022) to be upper bounded by $\xi(M, m)$ uniformly for all w. We divide both sides by m to obtain the theorem statement. □

In showing the simpler form of our bound also we use the following.

Proposition 4. *For any $m \geq 3$, $\xi(3, m) \leq 2m$.*

Proof. For $M = 3$ the upper bound in Theorem 5 evaluates to

$$\frac{1}{2} e^{\frac{1}{12m}} \left(1 + \frac{3}{\sqrt{m}} + \frac{6}{\pi m}\right) \cdot m.$$

The right hand part of this is a decreasing function of m and less than 2 for $m \geq M = 3$. □

Proof of Theorem 1. Set $M = 3$ and

$$\boldsymbol{U}_i(w) = \begin{bmatrix} \max(\tilde{\ell}(w, Z_{1:i}), 0) \\ \max(-\tilde{\ell}(w, Z_{1:i}), 0) \\ 1 - |\tilde{\ell}(w, Z_{1:i})| \end{bmatrix}$$

in Theorem 5, and bound $\xi(3, m)$ with Proposition 4. This gives the first part of the statement. For the second, we set $U_i = \mathbb{E}_{W\sim\rho_i^\star}[\ell(W, Z_i)]$ in Theorem 3 and note that $\mu_i = \mathbb{E}_{W\sim\rho_i^\star}\mathbb{E}_{Z\sim\mathcal{D}}[\ell(W, Z)]$ since $\rho_i^\star$ is independent of Z_i. The final part is obtained through a union bound of the two events. □

Proof of Theorem 2. The proof is the same as the first part of the proof of Theorem 1. □

Proof of Proposition 1. We know that $\operatorname{kl}([u, 0, 1-u] \| \boldsymbol{r}) \geq \operatorname{kl}(u \| r_1)$ by Adams et al. (2022, Proposition 9), with equality when $r_2 = 0$. Therefore we can set $r_2 = 0$ without making it any more difficult to satisfy the constraint $\operatorname{kl}([u, 0, 1-u] \| \boldsymbol{r}) \leq b$. In this case

$$\phi([u, 0, 1-u], b) = \sup\{r_1 : \operatorname{kl}(u \| r_1) \leq b\}$$

which is the definition of kl^{-1}. □

Proof of Proposition 3. Firstly, we recall (Adams et al., 2022, Proposition 9) that $\operatorname{kl}(\boldsymbol{u} \| \boldsymbol{v}) \geq \operatorname{kl}(u_i \| v_i)$ for any i. This immediately gives the first inequality upon inversion, if we note that $v_1 - v_2 \leq c$ for all $\operatorname{kl}(\boldsymbol{u} \| \boldsymbol{v}) \leq b$ implies that $\phi(\boldsymbol{u}, b) \leq c$. The first term is bounded with $v_1 \leq \operatorname{kl}^{-1}(u_1 \| b) \leq u_1 + \sqrt{2bu} + 2b$ as in the relaxation of Maurer's bound. Next we know that by Taylor's theorem, for any lower bound $0 \leq p < q \leq 1$, there exists $s \in [p, q]$ such that

$$\operatorname{kl}(q \| p) = \frac{(p-q)^2}{2s(1-s)} \geq \frac{(p-q)^2}{2q}.$$

Therefore

$$-v_2 \leq -u_2 + \sqrt{2bv_2}.$$

The proof is completed by summing these and applying the bound $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ for $a, b \geq 0$ (square both sides and subtract $a + b$ to reduce this to Young's inequality). □

5 EXPERIMENTS

In this section we empirically compare our bound to that of Maurer (2004) and the Unexpected Bernstein (Mhammedi et al., 2019), with a particular focus on the tightening arising through de-biasing by online estimators. For completeness, we give the exact statements of the bounds used in Appendix B, along with the procedure for calculating our bound.

We replicate the experimental setup of Mhammedi et al. (2019), looking which looks at classification with the 0-1 loss by logistic regression of UCI datasets.

The data space is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{0, 1\}$. Our hypotheses take the form $h_w(x) = \mathbf{1}_{\psi(w \cdot x) > \frac{1}{2}}$, where $\mathbf{1}$ is the

indicator function and $\psi(t) = 1/(1+e^{-t})$ is the standard logistic function. The 0-1 loss can be written as

$$\ell(w,(x,y)) = |y - \mathbf{1}_{\psi(w\cdot x)>\frac{1}{2}}|.$$

Specifically, we look at learning w by regularised logistic regression, which (for sample S and regularisation constant λ) outputs

$$\mathrm{LGR}_\lambda(S) = \operatorname*{argmin}_{w\in\mathcal{W}} \frac{\lambda\|w\|^2}{2} + \frac{1}{|S|}\sum_{(x,y)\in S}\sigma_w(x,y)$$

with

$$\sigma_w(x,y) = -y\log\psi(w\cdot x) - (1-y)\log(1-\psi(w\cdot x)).$$

This is solved empirically using the L-BFGS algorithm (Liu and Nocedal, 1989).

We set δ=0.05 and λ=0.01 on all datasets. In our bounds we choose posterior $\rho(S) = \mathcal{N}(\mathrm{LGR}_\lambda(S), \sigma^2 I)$, with $\sigma^2 \in \{1/2,\dots,1/2^J : J = \lceil \log_2 m\rceil\}$ chosen to minimise the bound being considered. For our prior we fix $\pi = \mathcal{N}(0,\sigma_0^2 I)$. Note that we are not using data-dependent priors as originally studied in Mhammedi et al. (2019), in order to isolate the effect of de-biasing; data-dependent PAC-Bayes priors are a rich topic in their own right.

The sequence of online estimators for our bound and the Unexpected Bernstein are chosen as the deterministic predictors outputted by $\mathrm{LGR}_\lambda(Z_{1:i-1})$; for computational reasons we update these only after every 150 examples, so that each new online estimator predicts the next 150 points. For the first 150 data points the online estimators are not yet effective so we simply choose them to have zero error, which does not change the bound on the loss of the online estimators.

The experiments use several UCI datasets, encoded and pre-processed using the same methods as Mhammedi et al. (2019). Specifically, we encode categorical variables in $0-1$ vectors (increasing the effective dimension of the feature space), remove any instances with missing features, and scale each feature to have values in $[-1,1]$. Experiments are repeated 20 times with different data shuffling and test-train allocation, and expectation with respect to Gaussian variables are evaluated using Monte Carlo estimates.

Discussion. Empirically we observe that our bound more effectively leverages the de-biasing of online estimators than the unexpected Bernstein, providing a tighter numerical guarantee in every case. On the smaller datasets it is somewhat weaker than Maurer's bound, but it is close to or better than it on the larger datasets. This arises because when the number of examples is very small, the online estimators are poor surrogates for the final posterior, and the de-biasing term is only weakly correlated with the loss. On the larger datasets, the de-biasing process is more effective and our bound is the tightest.

Dataset	Test	Maurer	UB	Ours
Haberman	0.273	**0.415**	0.583	0.501
Breast-C	0.037	**0.139**	0.208	0.164
Tictactoe	0.043	**0.214**	0.369	0.245
Banknote	0.050	**0.129**	0.192	0.136
kr-vs-kp	0.045	0.167	0.247	**0.164**
Spambase	0.169	0.324	0.501	**0.306**
Mushroom	0.003	**0.055**	0.082	0.056
Adult	0.170	0.234	0.384	**0.211**

Table 1: Test error of $\mathrm{LGR}_\lambda(S)$, and bounds for $\rho(S)$ with optimised σ as obtained on the UCI datasets listed. The datasets are ranked in order of size, from least examples to most. The bounds evaluated are Maurer's small-kl bound, the unexpected Bernstein bound and our Theorem 1, with the latter two using online estimators for de-biasing as described.

6 SUMMARY

In Theorem 1 we have provided a new PAC-Bayesian bound which can be used alongside an extension of excess losses. In particular, this extension of the excess loss is able to use information about the difficulty of examples in a pseudo-online fashion, as the learning algorithm passes over the dataset. This minimises the variance of our generalised excess loss. Our new bound is able to leverage this reduced variance to obtain tighter overall generalisation bounds and fast rates under broader settings.

By harnessing the power of online estimators and small-kl-based bounds in a new way, we have provided a new direction for numerical and theoretical improvements in PAC-Bayes bounds. Information about the difficulty of examples is most easily used for stable algorithms in our framework, which links nicely to further ideas like the complimentary use of data-dependent or distribution-dependent priors. In future work these same ideas could also be explored in an information-theoretic generalisation setting, such as the CMI setting of Steinke and Zakynthinou (2020); and extended to time-uniform bounds Haddouche and Guedj (2022).

Acknowledgements

F.B. acknowledges the support of the EPSRC grant EP/S021566/1. B.G. acknowledges partial support by the U.S. Army Research Laboratory, U.S. Army Research Office, U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1; B.G. also acknowledges partial support from the French National Agency for Research, grants ANR-18-CE40-0016-01 and ANR-18-CE23-0015-02.

References

Reuben Adams, John Shawe-Taylor, and Benjamin Guedj. Controlling confusion via generalisation bounds. *CoRR*, abs/2202.05560, 2022. URL https://arxiv.org/abs/2202.05560.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *CoRR*, abs/2110.11216, 2021. URL https://arxiv.org/abs/2110.11216.

Felix Biggs. A note on the efficient evaluation of PAC-Bayes bounds. *CoRR*, abs/2209.05188, 2022. doi: 10.48550/arXiv.2209.05188. URL https://doi.org/10.48550/arXiv.2209.05188.

Felix Biggs and Benjamin Guedj. Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23(10):1280, 2021. doi: 10.3390/e23101280. URL https://doi.org/10.3390/e23101280.

Felix Biggs and Benjamin Guedj. Non-vacuous generalisation bounds for shallow neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1963–1981. PMLR, 2022. URL https://proceedings.mlr.press/v162/biggs22a.html.

Felix Biggs, Valentina Zantedeschi, and Benjamin Guedj. On margins and generalisation for voting classifiers. In *NeurIPS*, 2022. doi: 10.48550/arXiv.2206.04607. URL https://doi.org/10.48550/arXiv.2206.04607.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics lecture notes-monograph series. Institute of Mathematical Statistics, 2007. ISBN 9780940600720. URL https://books.google.fr/books?id=acnaAAAAMAAJ.

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

Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems 31*, pages 8430–8441. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8063-data-dependent-pac-bayes-priors-via-differential-privacy.pdf.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 604–612. PMLR, 2021. URL http://proceedings.mlr.press/v130/karolina-dziugaite21a.html.

Andrew Y. K. Foong, Wessel P. Bruinsma, David R. Burt, and Richard E. Turner. How tight can PAC-Bayes be in the small data regime? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4093–4105, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/214cfbe603b7f9f9bc005d5f53f7a1d3-Abstract.html.

Andrew Y. K. Foong, Wessel P. Bruinsma, and David R. Burt. A note on the Chernoff bound for random variables in the unit interval. *CoRR*, abs/2205.07880, 2022. doi: 10.48550/arXiv.2205.07880. URL https://doi.org/10.48550/arXiv.2205.07880.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553419.

Benjamin Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, volume 33, 2019. URL https://arxiv.org/abs/1901.05353.

Maxime Haddouche and Benjamin Guedj. Pac-bayes with unbounded losses through supermartingales. *CoRR*, abs/2210.00928, 2022. doi: 10.48550/arXiv.2210.00928. URL https://doi.org/10.48550/arXiv.2210.00928.

Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, pages 713–721, 1956.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and Francois Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6872–6882. Curran Associates, Inc., 2019.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1-3):503–528, 1989. doi: 10.1007/

BF01589116. URL https://doi.org/10.1007/BF01589116.

Andrés R. Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian bounds for the weighted majority vote. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/386854131f58a556343e056f03626e00-Abstract.html.

Andreas Maurer. A note on the PAC-Bayesian theorem. *CoRR*, cs.LG/0411099, 2004. URL https://arxiv.org/abs/cs.LG/0411099.

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999.

David A. McAllester. Simplified PAC-Bayesian margin bounds. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, volume 2777 of *Lecture Notes in Computer Science*, pages 203–215. Springer, 2003. doi: 10.1007/978-3-540-45167-9_16.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes Un-Expected Bernstein inequality. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12180–12191, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3dea6b598a16b334a53145e78701fa87-Abstract.html.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes bounds with data dependent priors. *J. Mach. Learn. Res.*, 13:3507–3531, 2012. URL http://dl.acm.org/citation.cfm?id=2503353.

Maria Perez-Ortiz, Omar Rivasplata, Benjamin Guedj, Matthew Gleeson, Jingyu Zhang, John Shawe-Taylor, Miroslaw Bober, and Josef Kittler. Learning PAC-Bayes priors for probabilistic neural networks. 2021a. URL https://arxiv.org/abs/2109.10304.

Maria Perez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvari. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021b. URL http://jmlr.org/papers/v22/20-879.html.

Omar Rivasplata, Csaba Szepesvári, John Shawe-Taylor, Emilio Parrado-Hernández, and Shiliang Sun. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9234–9244, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/386854131f58a556343e056f03626e00-Abstract.html.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. In Nando de Freitas and Kevin P. Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, page 12. AUAI Press, 2012. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2341&proceeding_id=28.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. *CoRR*, abs/2001.09122, 2020. URL https://arxiv.org/abs/2001.09122.

Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 109–117, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/a97da629b098b75c294dffdc3e463904-Abstract.html.

Yi-Shan Wu and Yevgeny Seldin. Split-kl and PAC-Bayes-split-kl inequalities. In *NeurIPS*, volume abs/2206.00706, 2022. doi: 10.48550/arXiv.2206.00706. URL https://doi.org/10.48550/arXiv.2206.00706.

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning stochastic majority votes by minimizing a PAC-Bayes generalization bound. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 455–467, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0415740eaa4d9decbc8da001d3fd805f-Abstract.html.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: A PAC-Bayesian compression approach. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJgqqsAct7.

A ADDITIONAL PROOFS AND THEOREMS

A.1 PROOF OF PROPOSITION 2

Proof. We begin with the Donsker-Varadhan variational definition of the KL divergence for probability measures μ, ν on $\mathcal{A}$:

$$\mathrm{KL}(\mu,\nu) = \sup_{g:\mathcal{A}\to\mathbb{R}} \mathbb{E}_{A\sim\mu}[g(A)] - \log \mathbb{E}_{A\sim\nu}[\exp\circ g(A)].$$

Adapting to the case of the excess loss, $|\mathcal{A}| = 3$ and the probabilities of the possible events are measured by $\boldsymbol{q}, \boldsymbol{p} \in \Delta^3$. The set of real-valued functions on this space can be parameterised by a vector of values for each event, $\boldsymbol{r} \in \mathbb{R}^3$, so that the above equation is expressible as

$$\begin{aligned}
\mathrm{kl}(\boldsymbol{q},\boldsymbol{p}) &= \sup_{\boldsymbol{r}\in\mathbb{R}^3} \sum_{i=1}^3 q_i r_i - \log\left(\sum_{i=1}^3 p_i e^{r_i}\right) \\
&= \sup_{\boldsymbol{r}\in\mathbb{R}^3} q_1(r_1-r_3) + q_2(r_2-r_3) + r_3 - \log\left(e^{r_3}\left(p_1(e^{r_1-r_3}-1) + p_2(e^{r_2-r_3}-1)+1\right)\right) \\
&= \sup_{\boldsymbol{r}\in\mathbb{R}^3} q_1(r_1-r_3) + q_2(r_2-r_3) - \log\left(p_1(e^{r_1-r_3}-1) + p_2(e^{r_2-r_3}-1)+1\right).
\end{aligned}$$

Now we introduce the following reparamterisation of $r_1 - r_3 = \log(1-\eta), r_2 - r_3 = \log(1+\eta')$ so that

$$\mathrm{kl}(\boldsymbol{q},\boldsymbol{p}) = \sup_{\eta<1,\eta'>-1} -\log\left((-\eta p_1 + \eta' p_2 + 1)e^{-q_1\log(1-\eta) - q_2\log(1+\eta')}\right)$$

which can be re-arranged to give the following form very close to our final result

$$\sup_{\eta<1,\eta'>-1} \eta p_1 - \eta' p_2 - \left(1 - e^{q_1\log(1-\eta)+q_2\log(1+\eta')-\mathrm{kl}(\boldsymbol{q},\boldsymbol{p})}\right) = 0.$$

If we then relax the result by insisting that $\eta' = \eta \in (0,1)$, we find that

$$p_1 - p_2 \leq \inf_{\eta\in(0,1)} \frac{1}{\eta}\left(1 - e^{q_1\log(1-\eta)+q_2\log(1+\eta)-\mathrm{kl}(\boldsymbol{q},\boldsymbol{p})}\right)$$

which gives the first part of our result.

Rearranging terms,

$$\begin{aligned}
q_1\log(1-\eta) + q_2\log(1+\eta) &= q_1 - q_2 + q_1\left(\frac{-\log(1-\eta)-\eta}{\eta}\right) + q_2\left(\frac{-\log(1+\eta)+\eta}{\eta}\right) \\
&\leq q_1 - q_2 + (q_1+q_2)\left(\frac{-\log(1-\eta)-\eta}{\eta}\right) \\
&= q_1 - q_2 + c_\eta(q_1+q_2)
\end{aligned}$$

where it is easily verified that $-\log(1+\eta)+\eta \leq -\log(1-\eta)-\eta$ for $\eta > 0$.

Substitution of the definitions gives our proposition. □

A.2 PROOF OF THEOREM 3

Proof. We show that

$$\mathbb{P}\left(\bar{U} \geq \bar{\mu} + t\right) \leq e^{-m\cdot\mathrm{kl}(\bar{\mu}+t\|\bar{\mu})}. \tag{8}$$

From this, the proof of Theorem 4 in Biggs (2022) implies our theorem statement. By Markov's inequality and the convexity of $t \mapsto e^{ct}$,

$$\mathbb{P}\left(\bar{U} \geq \bar{\mu} + t\right) \;\leq\; \mathbb{P}\left(e^{m\lambda\bar{U}} \geq e^{m\lambda(\bar{\mu}+t)}\right)$$

$$
\begin{aligned}
&\leq e^{-m\lambda(\bar{\mu}+t)}\mathbb{E}\left[e^{m\lambda\bar{U}}\right] \\
&\leq e^{-m\lambda(\bar{\mu}+t)}\mathbb{E}\left[\prod_{i=1}^{m} e^{\lambda U_i}\right] \\
&\leq e^{-m\lambda(\bar{\mu}+t)}\mathbb{E}\left[\prod_{i=1}^{m}(1-U_i+U_i e^{\lambda})\right] \\
&\leq e^{-m\lambda(\bar{\mu}+t)}\mathbb{E}\left[\prod_{i=1}^{m}(1-\mu_i+\mu_i e^{\lambda})\right]
\end{aligned}
$$

where in the final step we have used the same telescoping property of conditional expectations as in the proof of Theorem 4. By the arithmetic-geometric mean inequality, the product term is upper bounded by

$$
\left(\frac{1}{m}\sum_{i=1}^{m}(1-\mu_i+\mu_i e^{\lambda})\right)^m = (1-\bar{\mu}+\bar{\mu}e^{\lambda})^m.
$$

Substitution shows that the probability above is upper bounded by

$$
\left(\frac{1-\bar{\mu}+\bar{\mu}e^{\lambda}}{e^{\lambda(\bar{\mu}+t)}}\right).
$$

Optimising this bound w.r.t. λ gives the form on Equation (8). □

A.3 PAC-BAYES UNEXPECTED BERNSTEIN WITH GENERALISED EXCESS LOSS

In this section we reproduce the following central result of Mhammedi et al. (2019) in the form used by our empirical comparison (which uses de-biasing but not informed priors).

Theorem 6 (PAC-Bayes unexpected Bernstein Excess Loss)**.** *For loss $\ell \in [0,1]$, for any fixed $\eta \in (0,1)$ and prior $\pi \in \mathcal{M}_1^+(\mathcal{W})$, with probability at least $1-\delta$ over the sample S simultaneously for any $\rho \in \mathcal{M}_1^+(\mathcal{W})$*

$$
\mathcal{E}(\rho) - \widehat{\mathcal{E}}(\rho) \leq c_\eta \widehat{V}(\rho) + \frac{\mathrm{KL}(\rho,\pi) + \log\frac{1}{\delta}}{m\eta}.
$$

Here $c_\eta = \frac{\eta+\log(1-\eta)}{\eta}$,

$$
\begin{aligned}
\mathcal{E}(\rho) &:= \mathcal{L}(\rho) - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_Z\mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z)], \\
\widehat{\mathcal{E}}(\rho) &:= \widehat{\mathcal{L}}(\rho) - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z)], \\
\widehat{V}(\rho) &:= \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^{m}|\ell(W,Z_i) - \mathbb{E}_{W'\sim\rho_i^\star}\ell(W',Z_i)|^2\right].
\end{aligned}
$$

We note that if the online estimators are fixed these quantities reduce to the standard excess risk terms. In order to prove this result, we first state and prove some intermediate results.

Proposition 5 (Unexpected Bernstein Lemma; Mhammedi et al., 2019, Lemma 13)**.** *Let $U \leq 1$ a.s.; then for any $\eta \in (0,1)$*

$$
\mathbb{E}e^{\eta(\mathbb{E}[U]-U-c_\eta U^2)} \leq 1
$$

where $c_\eta = \frac{\eta+\log(1-\eta)}{\eta}$.

Proof. For $t < 1$, define the decreasing function

$$
f(t) = \frac{\log(1-t)+t}{t^2}.
$$

Let $u \leq 1$ and $\eta \in (0,1)$, so that $u\eta \leq \eta < 1$. Since f is decreasing,

$$f(\eta) \leq f(u\eta) \implies \frac{\log(1-\eta)+\eta}{\eta^2} \leq \frac{\log(1-u\eta)+u\eta}{(u\eta)^2} \implies \exp(\eta c_\eta u^2 - \eta u) \leq 1 - u\eta.$$

Setting $u = U$ and taking the expectation, and using $1 - t \leq e^{-t}$,

$$\mathbb{E}\exp(\eta c_\eta U^2 - \eta U) \leq 1 - \mathbb{E}[U]\eta \leq e^{-\mathbb{E}[U]},$$

dividing through by the right hand side gives the result. □

We also give the following unexpected Bernstein counterpart of Theorem 5 which can be used to trivially prove the main result.

Theorem 7. *Let $U_1(w), \ldots, U_m(w)$ be a sequence of random bounded functions, valued in $[0,1]$, such that*

$$\mathbb{E}[U_i(w) | U_1(w), \ldots, U_{i-1}(w)] = \mu_i(w)$$

for $i = 1, \ldots, m$ and all $w \in \mathcal{W}$. Define

$$\widehat{U}(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m U_i(W)\right] \qquad \text{and} \qquad \bar{\mu}(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m \mu_i(W)\right].$$

For any fixed $\pi \in \mathcal{M}_1^+(\mathcal{W}), \delta \in (0,1), \eta \in (0,1)$, with probability at least $1-\delta$ (over all U_i), simultaneously for all $\rho \in \mathcal{M}_1^+(\mathcal{W})$,

$$\bar{\mu}(\rho) - \widehat{U}(\rho) \leq \frac{c_\eta}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho}|U_i(W)|^2 + \frac{\mathrm{KL}(\rho,\pi) + \log\frac{1}{\delta}}{m}$$

where $c_\eta = \frac{\eta + \log(1-\eta)}{\eta}$.

Proof. Firstly, we combine Proposition 5 with recursion of conditional expectations to find that

$$\begin{aligned}
&\mathbb{E}\exp\left(\eta\sum_{i=1}^m(\mu_i - U_i - c_\eta U_i^2)\right) \\
&\quad\leq \mathbb{E}\left[\prod_{i=1}^m \exp\left(\eta(\mu_i - U_i - c_\eta U_i^2)\right)\right] \\
&\quad\leq \mathbb{E}_{U_{1:m-1}}\left[\prod_{i=1}^{m-1}\exp\left(\eta(\mu_i - U_i - c_\eta U_i^2)\right)\cdot \mathbb{E}_{U_m}[\exp(\eta(\mu_m - U_m - c_\eta U_m^2))|U_{1:m-1}]\right] \\
&\quad\leq \mathbb{E}_{U_{1:m-1}}\left[\prod_{i=1}^{m-1}\exp(\eta(\mu_i - U_i - c_\eta U_i^2))\right] \\
&\quad\leq 1.
\end{aligned}$$

Next, as in the proof of Theorem 5, we combine this with Donsker-Varadhan, Markov's inequality and the independence of π from $U_{1:m}$ to find the following holds with probability at least $1-\delta$ for all ρ:

$$\begin{aligned}
&\mathbb{E}_{W\sim\rho}\left[\eta\sum_{i=1}^m(\mu_i(W) - U_i(W) - c_\eta U_i(W)^2)\right] - \mathrm{KL}(\rho,\pi) \\
&\qquad\leq \log\mathbb{E}_{W\sim\pi}\left[\eta\sum_{i=1}^m(\mu_i(W) - U_i(W) - c_\eta U_i(W)^2)\right] \\
&\qquad\leq \log\mathbb{E}_{U_{1:m}}\mathbb{E}_{W\sim\pi}\left[\eta\sum_{i=1}^m(\mu_i(W) - U_i(W) - c_\eta U_i(W)^2)\right]
\end{aligned}$$

$$\leq \log \frac{1}{\delta} \mathbb{E}_{W\sim\pi} \mathbb{E}_{U_{1:m}} \left[\eta \sum_{i=1}^{m} (\mu_i(W) - U_i(W) - c_\eta U_i(W)^2) \right]$$
$$\leq \log \frac{1}{\delta}.$$

Dividing both sides through by $m\eta$ gives the result. □

Proof of Theorem 6. In Theorem 7, set

$$U_i(w) = \ell(w, Z_i) - \mathbb{E}_{W\sim\rho_i^\star} \ell(W, Z_i)$$

for each i, which is bounded above by 1. □

A.4 RELAXATION OF SMALL KL

In the following, we prove a relaxation of the inverse small kl which leads to a form much more similar to the unexpected Bernstein, and is used later to motivate our experimental setup.

Proposition 6. *For* $0 \leq q < 1$ *and* $b > 0$*,*

$$\mathrm{kl}^{-1}(q\|b) < \inf_{\eta\in(0,1)} \left[(1 + c_\eta)q + \frac{b}{\eta} \right]$$

where $c_\eta := -\frac{\eta+\log(1-\eta)}{\eta}$.

In Germain et al. (2009, Proposition 2.1) it is proved that

Proposition 7. *For any* $0 \leq q \leq p < 1$*,*

$$\sup_{C>0} [C\Phi_C(p) - Cq] = \mathrm{kl}(q\|p)$$

where

$$\Phi_C(p) = -\frac{1}{C} \log(1 - p + pe^{-C}).$$

Proof of Proposition 6. For any $C > 0$, $C\Phi_C(p) - Cq \leq \mathrm{kl}(q\|p)$, and thus

$$\mathrm{kl}^{-1}(q\|b) = \sup\{p \in (0,1) : \mathrm{kl}(q\|p) \leq b\} \leq \sup\{p \in (0,1) : C\Phi_C(p) - Cq \leq b\} = \Phi_C^{-1}(q + b/c)$$

with the latter step following by the invertibility of Φ_C. Since $1 - e^{-t} \leq t$ with equality only at $t = 0$,

$$\Phi_C^{-1}(t) = \frac{1 - e^{-Ct}}{1 - e^{-C}} \leq \frac{Ct}{1 - e^{-C}}.$$

As $b > 0$ and $q \geq 0$, we have $q + b/c \neq 0$ and therefore

$$\Phi_C^{-1}(q + b/c) < \frac{Cq + b}{1 - e^{-C}}.$$

Introducing $\eta = 1 - e^{-C} \in (0, 1)$ (so that $C = -\log(1 - \eta)$),

$$\frac{Cq + b}{1 - e^{-C}} = \frac{-\log(1-\eta)}{\eta} q + \frac{b}{\eta} = (1 + c_\eta)q + \frac{b}{\eta}.$$

Chaining these results, and taking an infimum of both sides over the free variable $\eta \in (0, 1)$ completes the proof. □

B FULL BOUNDS USED IN EXPERIMENTS

In our experiments, we use the following bounds, obtained through Theorem 1 or (a slight refinement of) Theorem 6 with online estimators. In the unexpected Bernstein case, we combine the result with a grid over possible values of η, in the same way as the original paper. over possible values of η.

Theorem 8 (Generalisation Loss Bound)**.** *Fix $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{Z}), \pi \in \mathcal{M}_1^+(\pi), \delta \in (0,1)$, and $\ell : \mathcal{Z} \times \mathcal{W} \to [0,1]$. With probability at least $1-\delta$ over $Z_{1:m} = S \sim \mathcal{D}^m$, for a sequence of online estimators $\rho_i^\star \in \mathcal{M}_1^+(\mathcal{W})$ with $i = 1, \dots, m$, where each depenends only on the $i-1$ examples $Z_{1:i-1}$, and for any posterior $\rho \in \mathcal{M}_1^+(\mathcal{W})$, the following holds*

$$\mathcal{L}(\rho) \leq \phi\left(\begin{bmatrix} \widehat{\mathcal{E}}_+(\rho) \\ \widehat{\mathcal{E}}_-(\rho) \\ 1 - \widehat{\mathcal{E}}_+(\rho) - \widehat{\mathcal{E}}_-(\rho) \end{bmatrix}, \frac{\mathrm{KL}(\rho,\pi) + \log\frac{4m}{\delta}}{m}\right) + \mathrm{kl}^{-1}\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z_i)] \middle\| \frac{\log\frac{2}{\delta}}{m}\right)$$

with

$$\widehat{\mathcal{E}}_+(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m \max(\ell(W,Z_i) - \mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z_i)], 0)\right],$$
$$\widehat{\mathcal{E}}_-(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m \max(\mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z_i)] - \ell(W,Z_i), 0)\right].$$

Theorem 9 (Unexpected Bernstein for Generalisation Loss)**.** *Fix $\mathcal{D} \in \mathcal{M}_1^+(\mathcal{Z}), \pi \in \mathcal{M}_1^+(\pi), \delta \in (0,1)$, and $\ell : \mathcal{Z}\times\mathcal{W} \to [0,1]$. With probability at least $1-\delta$ over $Z_{1:m} = S \sim \mathcal{D}^m$, for a sequence of online estimators $\rho_i^\star \in \mathcal{M}_1^+(\mathcal{W})$ with $i = 1, \dots, m$, where each depenends only on the $i-1$ examples $Z_{1:i-1}$, and for any posterior $\rho \in \mathcal{M}_1^+(\mathcal{W})$, the following holds*

$$\mathcal{L}(\rho) \leq \widehat{\mathcal{L}}(\rho) + \inf_{\eta\in\mathcal{G}}\left[c_\eta \widehat{V}(\rho) + \frac{\mathrm{KL}(\rho,\pi) + \log\frac{2K}{\delta}}{m\eta}\right] + \mathrm{kl}^{-1}\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z_i)] \middle\| \frac{\log\frac{2}{\delta}}{m}\right)$$

where

$$\widehat{V}(\rho) := \mathbb{E}_{W\sim\rho}\left[\frac{1}{m}\sum_{i=1}^m |\ell(W,Z_i) - \mathbb{E}_{W'\sim\rho_i^\star}\ell(W',Z_i)|^2\right]$$

and

$$\mathcal{G} := \left\{\frac{1}{2}, \dots, \frac{1}{2^K} : K = \left\lceil \log_2\left(\frac{1}{2}\sqrt{\frac{n}{\log(1/\delta)}}\right)\right\rceil\right\}.$$

Proof. Combine Theorem 6 with the bound on $\mathcal{L}^\star$ in Theorem 1 and take a union bound over the grid $\mathcal{G}$. □

B.1 BOUNDING THE ONLINE ESTIMATOR LOSS

We note here that Mhammedi et al. (2019) originally used an alternative bound to go from the excess loss to the Generalisation risk. Instead of using the bound on $\mathcal{L}^\star$ in Theorem 1 as we do above, they used the bound

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_Z\mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z)] \leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z_i)] + \inf_{\eta\in\mathcal{G}}\left[\frac{c_\eta}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho_i^\star}|\ell(W,Z_i)|^2 + \frac{\mathrm{KL}(\rho,\pi) + \log\frac{|\mathcal{G}|}{\delta}}{m\eta}\right].$$

In the case of the 0-1 misclassification loss used in our experimental setup, where $\ell^2 = \ell$, this simplifies to the following:

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_Z\mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z)] \leq \inf_{\eta\in\mathcal{G}}\left[(1+c_\eta)\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho_i^\star}\ell(W,Z_i)\right) + \frac{\mathrm{KL}(\rho,\pi) + \log\frac{|\mathcal{G}|}{\delta}}{m\eta}\right].$$

As we find through Proposition 6, our Theorem 1 implies that

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_Z\mathbb{E}_{W\sim\rho_i^\star}[\ell(W,Z)] < \inf_{\eta\in(0,1)}\left[(1+c_\eta)\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{W\sim\rho_i^\star}\ell(W,Z_i)\right) + \frac{\mathrm{KL}(\rho,\pi) + \log\frac{1}{\delta}}{m\eta}\right]$$

which is strictly stronger (for example, our result holds simultaneously over all $\eta \in (0,1)$ with no grid size penalty, and even for the optimal η this bound is slacker). Therefore, the second part of our result Theorem 1 represents a significant contribution, that can be leveraged in combination with the original unexpected Bernstein bound to tighten it in the case of 0-1 losses (and it may also give tighter numerical bounds with some other loss functions also). We note that Theorem 1 can also easily be combined with the backwards-forwards dataset split used by Mhammedi et al. (2019).

In order to make the fairest empirical comparison between the effects of de-biasing on our bound versus the unexpected Bernstein, we therefore use our bound in the comparison.

B.2 CALCULATION OF INVERSE KL ϕ

Based directly on Proposition 11 in Adams et al. (2022), we give the following proposition, which can be used to calculate $\phi(\boldsymbol{u}, b)$.

Proposition 8. *Fix $\boldsymbol{u} \in \Delta_3$ and $b > 0$. Define the increasing function*

$$f_{\boldsymbol{u}}(s) := \log\left(u_1 + \frac{u_2}{1+2s} + \frac{u_3}{1+s}\right) + u_2 \log(1+2s) + u_3 \log(1+s)$$

and its inverse $s^\star := f_{\boldsymbol{u}}^{-1}(b)$. If $t := -\exp(-s^\star) - 1$, then

$$\phi(\boldsymbol{u}, b) = \frac{\frac{u_1}{t+1} - \frac{u_2}{t-1}}{\frac{u_1}{t+1} + \frac{u_2}{t-1} + \frac{u_3}{t}}.$$

Computationally, we can find the inverse $s^\star$ by a simple bisection-search or Newton's method. Our slight re-parameterisation (where we write f in terms of s instead of $t = -e^{-s} - 1$ as used by Adams et al., 2022) of the original result makes this calculation considerably more numerically stable.

We note as an aside that once we have calculated $s^\star$, we can also use it to find the gradients $\frac{\partial}{\partial u_i}\phi(\boldsymbol{u}, b)$ and $\frac{\partial}{\partial b}\phi(\boldsymbol{u}, b)$, which may be useful when directly optimising the bound as an objective.

C FURTHER EXPERIMENTAL DETAILS

Below we provide additional information about the datasets used and tabulated empirical results. Our code is available at https://github.com/biggs/tighter-pac-bayes-difficulty.

Dataset	Size	Test	Maurer	UB	Ours
Haberman	306	0.2726 ± 0.0388	0.4140 ± 0.0114	0.5829 ± 0.0176	0.5020 ± 0.0113
Breast-C	699	0.0371 ± 0.0133	0.1387 ± 0.0049	0.2079 ± 0.0070	0.1635 ± 0.0068
Tictactoe	958	0.0427 ± 0.0151	0.2148 ± 0.0056	0.3683 ± 0.0215	0.2456 ± 0.0069
Banknote	1372	0.0498 ± 0.0113	0.1292 ± 0.0033	0.1926 ± 0.0075	0.1359 ± 0.0038
kr-vs-kp	3196	0.0449 ± 0.0084	0.1670 ± 0.0023	0.2466 ± 0.0039	0.1633 ± 0.0029
Spambase	4601	0.1694 ± 0.0132	0.3238 ± 0.0027	0.5015 ± 0.0082	0.3054 ± 0.0032
Mushroom	8124	0.0026 ± 0.0013	0.0551 ± 0.0007	0.0820 ± 0.0015	0.0565 ± 0.0009
Adult	32561	0.1696 ± 0.0045	0.2341 ± 0.0013	0.3842 ± 0.0024	0.2108 ± 0.0014

Table 2: Test error of $\mathrm{LGR}_\lambda(S)$, and bounds for $\rho(S)$ with optimised σ as obtained on the UCI datasets listed. The datasets are ranked in order of size, from least examples to most. This size (listed) is the dataset size before the 20% test set is removed. The bounds evaluated are Maurer's small-kl bound, the unexpected Bernstein bound and our Theorem 1 as described in Appendix B, with the latter two using online estimators for de-biasing as in Section 5. Results are an average of 20 runs with standard errors provided.