
PAC-Bayes Un-Expected Bernstein Inequality

Zakaria Mhammedi

The Australian National University and Data61
zak.mhammedi@anu.edu.au

Peter D. Grünwald

CWI and Leiden University
pdg@cwi.nl

Benjamin Guedj

Inria and University College London
benjamin.guedj@inria.fr

Abstract

We present a new PAC-Bayesian generalization bound. Standard bounds contain a $\sqrt{L_n \cdot \text{KL}/n}$ complexity term which dominates unless L_n , the empirical error of the learning algorithm’s randomized predictions, vanishes. We manage to replace L_n by a term which vanishes in many more situations, essentially whenever the employed learning algorithm is sufficiently stable on the dataset at hand. Our new bound consistently beats state-of-the-art bounds both on a toy example and on UCI datasets (with large enough n). Theoretically, unlike existing bounds, our new bound can be expected to converge to 0 faster whenever a Bernstein/Tsybakov condition holds, thus connecting PAC-Bayesian generalization and *excess risk* bounds — for the latter it has long been known that faster convergence can be obtained under Bernstein conditions. Our main technical tool is a new concentration inequality which is like Bernstein’s but with X^2 taken outside its expectation.

1 Introduction

PAC-Bayesian generalization bounds [1, 7, 8, 14, 15, 17, 24–26] have recently obtained renewed interest within the context of deep neural networks [12, 30, 36]. In particular, Zhou et al. [36] and Dziugaite and Roy [12] showed that, by extending an idea due to Langford and Caruana [19], one can obtain nontrivial (but still not very strong) generalization bounds on real-world datasets such as MNIST and ImageNet. Since using alternative methods, nontrivial generalization bounds are even harder to get, there remains a strong interest in improved PAC-Bayesian bounds. In this paper, we provide a considerably improved bound whenever the employed learning algorithm is sufficiently *stable* on the given data.

Standard bounds all have an order $\sqrt{L_n \cdot \text{COMP}_n/n}$ term on the right, where COMP_n represents model complexity in the form of a Kullback-Leibler divergence between a prior and a posterior, and L_n is the *posterior expected loss* on the training sample. The latter only vanishes if there is a sufficiently large neighborhood around the ‘center’ of the posterior at which the training error is 0. In the two papers [12, 36] mentioned above, this is not the case. For example, the various deep net experiments reported by Dziugaite et al. [12, Table 1] with $n = 150000$ all have L_n around 0.03, so that $\sqrt{\text{COMP}_n/n}$ is multiplied by a non-negligible $\sqrt{0.03} \approx 0.17$. Furthermore, they have COMP_n increasing substantially with n , making $\sqrt{L_n \cdot \text{COMP}_n/n}$ converge to 0 at rate slower than $1/\sqrt{n}$.

In this paper, we provide a bound (Theorem 1) with L_n replaced by a second-order term V_n — a term which will go to 0 in many cases in which L_n does not. This can be viewed as an extension of an earlier second-order approach by Tolstikhin and Seldin [34] (TS from now on): they also replace L_n , but by a term that, while usually smaller than L_n , will tend to be larger than our V_n . Specifically,

as they write, in classification settings (our primary interest), their replacement is not much smaller than L_n itself. Instead our V_n can be very close to 0 in classification even when L_n is large. While the TS bound is based on an ‘empirical’ Bernstein inequality due to [23], our bound is based on a different modification of Bernstein’s moment inequality in which the occurrence of X^2 is taken outside of its expectation. We call this result, which is of independent interest, the *un*-expected Bernstein inequality — see Lemma 10.

The term V_n in our bound goes to 0 — and our bound improves on existing bounds — whenever the employed learning algorithm is relatively stable on the given data; for example, if the predictor learned on an initial segment (say, 50%) of the dataset performs similarly (*i.e.* assigns similar losses to the same samples) to the predictor based on the full data. This improvement is reflected in our experiments where, except for very small sample sizes, we consistently outperform existing bounds both on a toy classification problem with label noise and on standard UCI datasets [11]. Of course, the importance of stability for generalization has been recognized before in landmark papers such as [6, 29, 33]. However, the data-dependent stability notion ‘ V_n ’ occurring in our bound seems very different from any of the notions discussed in those papers.

Theoretically, a further contribution is that we connect our PAC-Bayesian generalization bound to *excess risk bounds*: we show that (Theorem 4) our generalization bound can be of comparable size to excess risk bounds up to an irreducible *complexity-free* term that is independent of model complexity. The excess risk bound that can be attained for any given problem depends both on the complexity of the set of predictors \mathcal{H} and on the inherent ‘easiness’ of the problem. The latter is often measured in terms of the exponent $\beta \in [0, 1]$ of the *Bernstein condition* that holds for the given problem [5, 13, 16], which generalizes the exponent in the celebrated *Tsybakov margin condition* [4, 35] (this has been a main topic in two recent NeurIPS workshops on ‘learning from easy data’). The larger β , the faster the excess risk converges. In Section 5, we essentially show that the rate at which the (often dominating) $\sqrt{V_n \cdot \text{COMP}_n/n}$ term goes to 0 can also be bounded by a quantity that gets smaller as β gets larger. In contrast, previous PAC-Bayesian bounds do not have such a property.

Contents. In Section 2, we introduce the problem setting and provide a first, simplified version of our theorem. Section 3 gives our main bound. Experiments are presented in Section 4, followed by theoretical motivation in Section 5. The proof of our main bound is provided in Section 6, where we first present the convenient ESI language for expressing stochastic inequalities, and (our main tool) the unexpected Bernstein inequality in Lemma 10. The paper ends with an outlook for future work.

2 Problem Setting, Background, and Simplified Version of Our Bound

Setting and Notation. Let Z_1, \dots, Z_n be i.i.d. random variables in some set \mathcal{Z} , with $Z_i \sim \mathbf{D}$, for $i \in [n]$. Let \mathcal{H} be a hypothesis set and $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, b]$, $b > 0$, be a bounded loss function such that $\ell_h(Z) := \ell(h, Z)$ denotes the loss that hypothesis h makes on Z . We call any such tuple $(\mathbf{D}, \ell, \mathcal{H})$ a *learning problem*. For a given hypothesis $h \in \mathcal{H}$, we denote its *risk* (expected loss on a test sample of size 1) by $L(h) := \mathbb{E}_{Z \sim \mathbf{D}} [\ell_h(Z)]$ and its empirical error by $L_n(h) := \frac{1}{n} \sum_{i=1}^n \ell_h(Z_i)$. For any distribution P on \mathcal{H} , we write $L(P) := \mathbb{E}_{h \sim P} [L(h)]$ and $L_n(P) := \mathbb{E}_{h \sim P} [L_n(h)]$.

For any $m \in [n]$ and any variables Z_1, \dots, Z_n in \mathcal{Z} , we denote $Z_{\leq m} := (Z_1, \dots, Z_m)$ and $Z_{< m} := Z_{\leq m-1}$, with the convention that $Z_{\leq 0} = \emptyset$. We follow a similar rule for $Z_{\geq m}$ and $Z_{> m}$. As is customary in PAC-Bayesian works, a *learning algorithm* is a (computable) function $P : \bigcup_{i=1}^n \mathcal{Z}_i \rightarrow \Delta(\mathcal{H})$ that, upon observing input $Z_{\leq n} \in \mathcal{Z}^n$, outputs a ‘posterior’ distribution $P(Z_{\leq n})(\cdot)$ on \mathcal{H} . The posterior could be a Gibbs or a generalized-Bayesian posterior but also other algorithms. When no confusion can arise, we will abbreviate $P(Z_{\leq n})$ to P_n , and denote P_0 any ‘prior’ distribution, *i.e.*, a distribution on \mathcal{H} which has to be specified in advance, before seeing the data. Finally, we denote the Kullback-Leibler divergence between P_n and P_0 by $\text{KL}(P_n \| P_0)$.

Comparing Bounds. Both existing state-of-the-art PAC-Bayes bounds and ours essentially take the following form: there exists constants $\mathcal{P}, \mathcal{A}, \mathcal{C} \geq 0$, and a function $s_{\delta, n}$ (s for small), logarithmic in $1/\delta$ and n , such that $\forall \delta > 0$, with probability at least $1 - \delta$ over the sample Z_1, \dots, Z_n , it holds that,

$$L(P_n) - L_n(P_n) \leq \mathcal{P} \cdot \sqrt{\frac{R_n \cdot (\text{COMP}_n + s_{\delta, n})}{n}} + \mathcal{A} \cdot \frac{\text{COMP}_n + s_{\delta, n}}{n} + \mathcal{C} \cdot \sqrt{\frac{R'_n \cdot s_{\delta, n}}{n}}, \quad (1)$$

where $R_n, R'_n \geq 0$ are sample-dependent quantities which may differ from one bound to another. Existing classical bounds that after slight relaxations take on this form are due to Langford and Seeger [20, 32], Catoni [9], Maurer [22], and Tolstikhin and Seldin (TS) [34] (see the latter for a nice overview). In all these cases, $\text{COMP}_n = \text{KL}(P_n \| P_0)$, $R'_n = 0$, and — except for the TS bound — $R_n = L_n(P_n)$. For the TS bound, R_n is equal to empirical loss variance. Our bound in Theorem 1 also fits (1) (after a relaxation), but with considerably different choices for COMP_n , R'_n , and R_n .

Of special relevance in our experiments is the bound due to Maurer [22], which as noted by TS [34] tightens the PAC-Bayes-kl inequality due to Seeger [31], and is one of the tightest known generalization bounds in the literature. It can be stated as follows: for $\delta \in]0, 1[$, $n \geq 8$, and any learning algorithm P , with probability at least $1 - \delta$,

$$\text{kl}(L(P_n), L_n(P_n)) \leq \frac{\text{KL}(P_n \| P_0) + \ln \frac{2\sqrt{n}}{\delta}}{n}, \quad (2)$$

where kl is the binary Kullback-Leibler divergence. Applying the inequality $p \leq q + \sqrt{2q \text{kl}(p \| q)} + 2 \text{kl}(p \| q)$ to (2) yields a bound of the form (1) (see [34] for more details). Note also that using Pinsker's inequality together with (2) implies McAllester's classical PAC-Bayesian bound [24].

Corollary of Theorem 1 below. For any deterministic estimator $\hat{h} : \cup_{i=1}^n \mathcal{Z}_i \rightarrow \mathcal{H}$ (such as ERM or a SGD iterate), there exists $\mathcal{P}, \mathcal{A}, \mathcal{C} > 0$, such that (1) holds with probability at least $1 - \delta$, with

$$\text{COMP}_n = \text{KL}(P_n \| P(Z_{\leq m})) + \text{KL}(P_n \| P(Z_{> m})).$$

$$R_n = V_n := \frac{1}{n} \mathbb{E}_{h \sim P_n} \left[\sum_{i=1}^m \left(\ell_h(Z_i) - \ell_{\hat{h}(Z_{> m})}(Z_i) \right)^2 + \sum_{j=m+1}^n \left(\ell_h(Z_j) - \ell_{\hat{h}(Z_{\leq m})}(Z_j) \right)^2 \right],$$

and R'_n at most $O(b^2)$, where $b > 0$ is an upper-bound on the loss ℓ . Like in TS's and Catoni's bound, but unlike McAllester's and Maurer's, our $s_{\delta, n}$ grows as $\ln((\ln n)/\delta)$. A larger difference is that our complexity term is a sum of two KL divergences, in which the prior is 'informed' — when $m = n/2$, it is really the posterior based on half the sample. Our experiments confirm that this tends to be much smaller than $\text{KL}(P_n \| P_0)$. Note that one can make the existing bounds conceptually closer to ours by learning a classifier based on the first half of the data, and using a prior 'centered' at that classifier to obtain a PAC-Bayes bound on the second half — an idea pioneered by [2]. In additional experiments reported in Appendix A, we found that while this improves existing bounds in some cases it also worsens them in others, and the conclusions of Section 4 (the experiments) remain the same (see Appendix A).

In light of the preceding observation, we regard the fact that we have $R_n = V_n$ in our bound instead of $R_n = L_n(P_n)$ as the more fundamental difference to other approaches. Only TS [34] have a R_n that is somewhat reminiscent of ours: in their case $R_n = \mathbb{E}_{h \sim P_n} [\sum_{i=1}^n (\ell_h(Z_i) - L_n(h))^2] / (n - 1)$ is the empirical loss variance. The crucial difference to our V_n is that the empirical loss variance cannot be close to 0 unless a sizeable P_n -posterior region of h has empirical error almost constant on most data instances. For classification with 0-1 loss, this is a strong condition since the empirical loss variance is equal to $nL_n(P_n)(1 - L_n(P_n)) / (n - 1)$, which is only close to 0 if $L_n(P_n)$ is itself close to 0 or 1. In contrast, our V_n can go to zero 0 even if the empirical error and variance do not. This can be witnessed in our experiments in Section 4. In Section 5, we argue more formally that under a Bernstein condition, the $\sqrt{V_n \cdot \text{COMP}_n / n}$ term can be much smaller than $\sqrt{\text{COMP}_n / n}$. Note, finally, that the term V_n has 2-fold cross-validation flavor, but in contrast to a cross-validation error, for V_n to be small, it is sufficient that the losses are *similar*, not that they are small.

The price we pay for all this (and that does not show up in other bounds) is the right-most, irreducible remainder term in (1) of order at most b/\sqrt{n} (up to log-factors). Note, however, that this term is decoupled from the complexity COMP_n , and thus it is not affected by COMP_n growing with n . We call this an 'irreducible' term, because, by the central limit theorem, a $O(1/\sqrt{n})$ term is unavoidable in any PAC-Bayesian bound: this is the case even if $\mathcal{H} = \{h\}$ is a singleton and there is no learning, unless $\ell_h(Z)$ has zero-variance.

3 Main Bound

We now present our main result in its most general form. Let $\vartheta(\eta) := (-\ln(1 - \eta) - \eta)/\eta^2$ and $c_\eta := \eta \cdot \vartheta(\eta b)$, for $\eta \in]0, 1/b[$, where $b > 0$ is an upper-bound on the loss ℓ .

Theorem 1. [Main Theorem] Let Z_1, \dots, Z_n be i.i.d. with $Z_i \sim \mathbf{D}$, for $i \in [n]$. Let $m \in \{0, \dots, n\}$ and π be any distribution with support on a finite or countable grid $\mathcal{G} \subset]0, 1/b[$. For any $\delta \in]0, 1[$, and any learning algorithms $P, Q : \bigcup_{i=1}^n \mathcal{Z}_i \rightarrow \Delta(\mathcal{H})$, we have,

$$L(P_n) \leq L_n(P_n) + \inf_{\eta \in \mathcal{G}} \left\{ c_\eta \cdot V_n + \frac{\text{COMP}_n + 2 \ln \frac{1}{\delta \cdot \pi(\eta)}}{\eta \cdot n} \right\} + \inf_{\nu \in \mathcal{G}} \left\{ c_\nu \cdot V'_n + \frac{\ln \frac{1}{\delta \cdot \pi(\nu)}}{\nu \cdot n} \right\}, \quad (3)$$

with probability at least $1 - \delta$, where COMP_n, V'_n , and V_n are the random variables defined by:

$$\begin{aligned} \text{COMP}_n &:= \text{KL}(P_n \| P(Z_{\leq m})) + \text{KL}(P_n \| P(Z_{> m})), \\ V'_n &:= \frac{1}{n} \cdot \left(\sum_{i=1}^m \mathbb{E}_{h' \sim Q(Z_{> i})} [\ell_{h'}(Z_i)^2] + \sum_{j=m+1}^n \mathbb{E}_{h' \sim Q(Z_{< j})} [\ell_{h'}(Z_j)^2] \right), \\ V_n &:= \frac{1}{n} \mathbb{E}_{h \sim P_n} \left[\sum_{i=1}^m (\ell_h(Z_i) - \mathbb{E}_{h' \sim Q(Z_{> i})} [\ell_{h'}(Z_i)])^2 + \sum_{j=m+1}^n (\ell_h(Z_j) - \mathbb{E}_{h' \sim Q(Z_{< j})} [\ell_{h'}(Z_j)])^2 \right]. \end{aligned}$$

While the result holds for all $0 \leq m \leq n$, in the remainder of this paper, we assume for simplicity that n is even and that $m = n/2$. We will also be using the grid \mathcal{G} and distribution π defined by

$$\mathcal{G} := \left\{ \frac{1}{2b}, \dots, \frac{1}{2^{Kb}} : K := \left\lceil \log_2 \left(\frac{1}{2} \sqrt{\frac{n}{\ln \frac{1}{\delta}}} \right) \right\rceil \right\}, \text{ and } \pi \equiv \text{uniform distribution over } \mathcal{G}. \quad (4)$$

Roughly speaking, this choice of \mathcal{G} ensures that the infima in η and ν in (3) are attained within $[\min \mathcal{G}, \max \mathcal{G}]$. Using the relaxation $c_\eta \leq \eta/2 + \eta^2 11b/20$, for $\eta \leq 1/(2b)$, in (3) and tuning η and ν within the grid \mathcal{G} defined in (4) leads to a bound of the form (1). Furthermore, we see that the expression of V_n in the simplified version of Theorem 1 given in the previous section now follows when Q is chosen such that, for $1 \leq i \leq m < j \leq n$, $Q(Z_{> i}) \equiv \delta(\hat{h}(Z_{> m}))$ and $Q(Z_{< j}) \equiv \delta(\hat{h}(Z_{\leq m}))$, for some deterministic estimator \hat{h} , where $\delta(h)(\cdot)$ denotes the Dirac distribution at $h \in \mathcal{H}$. It is clear that Theorem 1 is considerably more general than its corollary above: when predicting the j -th point Z_j ($j > m$) in the second term of V_n , we could use an estimator that does not only depend on Z_1, \dots, Z_m but also on part of the second sample, namely Z_{m+1}, \dots, Z_{j-1} , and analogously for predicting Z_i ($i \leq m$) in the first term. We can thus base our bound on a sum of errors achieved by *online* estimators that converge to the final $\hat{h}(Z_{\leq n})$ based on the full data. Doing this would likely improve our bounds, but is computationally demanding, and so we did not try it in our experiments.

Remark 2. (Useful for Section 5 below) Though this may deteriorate the bound in practice, Theorem 1 allows choosing a learning algorithm P such that for $1 < m < n$, $P(Z_{\leq m}) \equiv P(Z_{> m}) \equiv P_0$ (hence only P_n ‘truly’ learns); this results in $\text{COMP}_n = 2\text{KL}(P_n \| P_0)$ — the bound is otherwise unchanged.

4 Experiments

In this section, we experimentally compare our bound in Theorem 1 to that of TS [34], Catoni [8, Theorem 1.2.8] (with $\alpha = 2$), and Maurer in (2)¹. For the latter, given $L_n(P_n) \in [0, 1[$ and the RHS of (2), we solve for an upper bound of $L(P_n)$ by ‘inverting’ the kl. We note that TS [34] do not claim that their bound is better than Maurer’s in classification (in fact, they do better in other settings).

Setting. We consider both synthetic and real-world datasets for binary classification, and we evaluate bounds using the 0-1 loss. In particular, the data space \mathcal{Z} is $\mathcal{X} \times \mathcal{Y} := \mathbb{R}^d \times \{0, 1\}$, where $d \in \mathbb{N}$ is the dimension of the feature space. In this case, the hypothesis set \mathcal{H} is also \mathbb{R}^d , and the error associated with $h \in \mathcal{H}$ on a sample $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$ is given by $\ell_h(Z) = |Y - \mathbb{1}\{\phi(h^\top X) > 1/2\}|$, where $\phi(w) := 1/(1 + e^{-w})$, $w \in \mathbb{R}$. We learn our hypotheses using *regularized logistic regression*; given a sample $S = (Z_p, \dots, Z_q)$, with $1 \leq p < q \leq n$, we compute

$$\hat{h}(S) := \arg \min_{h \in \mathcal{H}} \frac{\lambda \|h\|^2}{2} + \frac{1}{q-p+1} \sum_{i=p}^q Y_i \cdot \ln \phi(h^\top X_i) + (1 - Y_i) \cdot \ln(1 - \phi(h^\top X_i)). \quad (5)$$

¹Find code at <https://github.com/bguedj/PAC-Bayesian-Un-Expected-Bernstein-Inequality>.

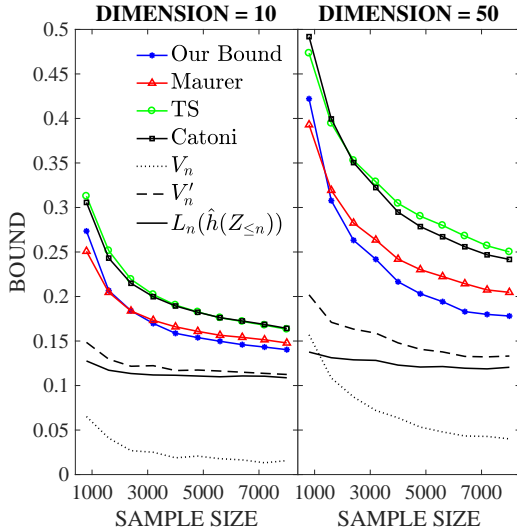


Figure 1: Results for the synthetic data.

UCI	d	Test err.	Our bound	Maurer bound
Dataset		of \hat{h}		
Haberman	3	0.272	0.521	0.411
($n=244$)				
Breast-C.	9	0.068	0.185	0.159
($n=560$)				
TicTacToe	27	0.046	0.191	0.216
($n=766$)				
Banknote	4	0.058	0.125	0.136
($n=1098$)				
kr-vs-kp	73	0.044	0.108	0.165
($n=2556$)				
Spambase	57	0.173	0.293	0.312
($n=3680$)				
Mushroom	116	0.002	0.018	0.055
($n=6500$)				
Adult	108	0.168	0.195	0.234
($n=24130$)				

Table 1: Results for the UCI datasets.

For $Z_{\leq n} \in \mathcal{Z}^n$, and $1 \leq i \leq m < j \leq n$, we choose algorithm Q in Theorem 1 such that

$$Q(Z_{>i}) \equiv \delta(\hat{h}(Z_{>m})) \quad \text{and} \quad Q(Z_{<j}) \equiv \delta(\hat{h}(Z_{\leq m})).$$

Given a sample $S \neq \emptyset$, we set the ‘posterior’ $P(S)$ to be a Gaussian centered at $\hat{h}(S)$ with variance $\sigma^2 > 0$: that is, $P(S) \equiv \mathcal{N}(\hat{h}(S), \sigma^2 I_d)$. The prior distribution is set to $P_0 \equiv \mathcal{N}(0, \sigma_0^2 I_d)$, for $\sigma_0 > 0$.

Parameters. We set $\delta = 0.05$. For all datasets, we use $\lambda = 0.01$, and (approximately) solve (5) using the BFGS algorithm. For each bound, we pick the $\sigma^2 \in \{1/2, \dots, 1/2^J : J := \lceil \log_2 n \rceil\}$ which minimizes it on the given data (with n instances). In order for the bounds to still hold with probability at least $1 - \delta$, we replace δ on the RHS of each bound by $\delta / \lceil \log_2 n \rceil$ (this follows from the application of the union bound). We choose the prior variance such that $\sigma_0^2 = 1/2$ (this was the best value on average for the bounds we compare against). We choose the grid \mathcal{G} in Theorem 1 as in (4). Finally, we approximate Gaussian expectations using *Monte Carlo* sampling.

Synthetic data. We generate synthetic data for $d = \{10, 50\}$ and sample sizes between 800 and 8000. For a given sample size n , we 1) draw X_1, \dots, X_n [resp. $\varepsilon_1, \dots, \varepsilon_n$] identically and independently from the multivariate-Gaussian distribution $\mathcal{N}(0, I_d)$ [resp. the Bernoulli distribution $\mathcal{B}(0.9)$]; and 2) we set $Y_i = \mathbb{1}\{\phi(h_*^\top X_i) > 1/2\} \cdot \varepsilon_i$, for $i \in [n]$, where $h_* \in \mathbb{R}^d$ is the vector constructed from the first d digits of π . For example, if $d = 10$, then $h_* = (3, 1, 4, 1, 5, 9, 2, 6, 5, 3)^\top$. Figure 1 shows the results averaged over 10 independent runs for each sample size.

UCI datasets. For the second experiment, we use several UCI datasets. These are listed in Table 1 (where Breast-C. stands for Breast Cancer). We encode categorical variables in appropriate 0-1 vectors. This effectively increases the dimension of the input space (this is reported as d in Table 1). After removing any rows (*i.e.* instances) containing missing features and performing the encoding, the input data is scaled such that every column has values between -1 and 1. We used a 5-fold train-test split (n in Table 1 is the training set size), and the results in Table 1 are averages over 5 runs. We only compare with Maurer’s bound since other bounds were worse than Maurer’s and ours on all datasets.

Discussion. As the dimension d of the input space increases, the complexity $\text{KL}(P_n \| P_0)$ — and thus, all the PAC-Bayes bounds discussed in this paper — get larger. Our bound suffers less from this increase in d , since for a large enough sample size n , the term V_n is small enough (see Figure 1) to absorb any increase in the complexity. In fact, for large enough n , the irreducible (complexity-free) term involving V'_n in our bound becomes the dominant one. This, combined with the fact that for the 0-1 loss, $V'_n \approx L_n(P_n)$ for large enough n (see Figure 1), makes our bound tighter than others.

Adding a regularization term in the objective (5) is important as it stabilizes $\hat{h}(Z_{<m})$ and $\hat{h}(Z_{\geq m})$; a similar effect is achieved with methods like gradient descent as they essentially have a ‘built-in’ regularization. For very small sample sizes, the regularization in (5) may not be enough to ensure that $\hat{h}(Z_{<m})$ and $\hat{h}(Z_{\geq m})$ are close to $\hat{h}(Z_{\leq n})$, in which case V_n need not be necessarily small. In particular, this is the case for the Haberman and the breast cancer datasets where the advantage of our bound is not fully leveraged, and Maurer’s bound is smaller

5 Theoretical Motivation of the Bound

In this section, we study the behavior of our bound (3) under a Bernstein condition:

Definition 3. [Bernstein Condition (BC)] *The learning problem $(\mathbf{D}, \ell, \mathcal{H})$ satisfies the (β, B) -Bernstein condition, for $\beta \in [0, 1]$ and $B > 0$, if for all $h \in \mathcal{H}$,*

$$\mathbb{E}_{Z \sim \mathbf{D}} \left[(\ell_h(Z) - \ell_{h_*}(Z))^2 \right] \leq B \cdot \mathbb{E}_{Z \sim \mathbf{D}} [\ell_h(Z) - \ell_{h_*}(Z)]^\beta,$$

where $h_* := \arg \inf_{h \in \mathcal{H}} \mathbb{E}_{Z \sim \mathbf{D}} [\ell_h(Z)]$ is the risk minimizer within the class \mathcal{H} .

The Bernstein condition [3–5, 13, 18] essentially characterizes the ‘easiness’ of the learning problem; it implies that the variance in the excess loss random variable $\ell_h(Z) - \ell_{h_*}(Z)$ gets smaller the closer the risk of hypothesis $h \in \mathcal{H}$ gets to that of the risk minimizer h_* . For bounded loss functions, the BC with $\beta = 0$ always holds. The BC with $\beta = 1$ (the ‘easiest’ learning setting) is also known as the *Massart noise condition* [21]; it holds in our experiment with synthetic data in Section 4, and also, e.g., whenever \mathcal{H} is convex and $h \mapsto \ell_h(z)$ is exp-concave, for all $z \in \mathcal{Z}$ [13, 27]. For more examples of learning settings where a BC holds see [18, Section 3].

Our aim in this section is to give an upper-bound on the infimum term involving V_n in (3), under a BC, in terms of the complexity COMP_n and the excess risks $\bar{L}(P_n)$, $\bar{L}(Q(Z_{>m}))$, and $\bar{L}(Q(Z_{\leq m}))$, where for a distribution $P \in \Delta(\mathcal{H})$, the excess risk is defined by

$$\bar{L}(P) := \mathbb{E}_{h \sim P} [\mathbb{E}_{Z \sim \mathbf{D}} [\ell_h(Z)]] - \mathbb{E}_{Z \sim \mathbf{D}} [\ell_{h_*}(Z)].$$

In the next theorem, we denote $Q_{\leq m} := Q(Z_{\leq m})$ and $Q_{> m} := Q(Z_{> m})$, for $m \in [n]$. To simplify the presentation further (and for consistency with Section 4), we assume that Q is chosen such that

$$Q(Z_{>i}) = Q_{>m}, \text{ for } 1 \leq i \leq m, \quad \text{and} \quad Q(Z_{<j}) = Q_{\leq m}, \text{ for } m < j \leq n. \quad (6)$$

Theorem 4. *Let \mathcal{G} and π be as in (4), $\delta \in]0, 1[$, and $s_{\delta,n} = 2 \ln \frac{1}{\delta \cdot \pi(\eta)}$, $\eta \in \mathcal{G}$. If the (β, B) -Bernstein condition holds with $\beta \in [0, 1]$ and $B > 0$, then for any learning algorithms P and Q (with Q satisfying (6)), there exists a $\mathcal{C} > 0$, such that $\forall n \geq 1$ and $m = n/2$, with probability at least $1 - \delta$,*

$$\begin{aligned} \frac{1}{\mathcal{C}} \cdot \inf_{\eta \in \mathcal{G}} \left\{ c_\eta \cdot V_n + \frac{\text{COMP}_n + s_{\delta,n}}{\eta \cdot n} \right\} &\leq \bar{L}(P_n) + \bar{L}(Q_{\leq m}) + \bar{L}(Q_{>m}) \\ &+ \left(\frac{\text{COMP}_n + s_{\delta,n}}{n} \right)^{\frac{1}{2-\beta}} + \frac{\text{COMP}_n + s_{\delta,n}}{n}. \end{aligned} \quad (7)$$

In addition to the ‘ESI’ tools provided in Section 6 and Lemma 10, the proof of Theorem 4, presented in Appendix C, also uses an ‘ESI version’ of the Bernstein condition due to [18].

First note that the only terms in our main bound (3), other than the infimum on the LHS of (7), are the empirical error $L_n(P_n)$ and a $\tilde{O}(1/\sqrt{n})$ -complexity-free term which is typically smaller than $\sqrt{\text{KL}(P_n \| P_0)/n}$ (e.g. when the dimension of \mathcal{H} is large enough). The latter term is often the dominating one in other PAC-Bayesian bounds when $\liminf_{n \rightarrow \infty} L_n(P_n) > 0$.

Now consider the remaining term in our main bound, which matches the infimum term on the LHS of (7), and let us choose algorithm P as per Remark 2, so that $\text{COMP}_n = 2\text{KL}(P_n \| P_0)$. Suppose that, with high probability (w.h.p), $\text{KL}(P_n \| P_0)/n$ converges to 0 for $n \rightarrow \infty$ (otherwise no PAC-Bayesian bound would converge to 0), then $(\text{COMP}_n/n)^{1/(2-\beta)} + \text{COMP}_n/n$ — essentially the sum of the last two terms on the RHS of (7) — converges to 0 at a faster rate than $\sqrt{\text{KL}(P_n \| P_0)/n}$ w.h.p. for $\beta > 0$, and at equal rate for $\beta = 0$. Thus, in light of Theorem 4, to argue that our bound can be

better than others (still when $\liminf_{n \rightarrow \infty} L_n(P_n) > 0$), it remains to show that there exist algorithms P and Q for which the sum of the excess risks on the RHS of (7) is smaller than $\sqrt{\text{KL}(P_n \| P_0)}/n$.

One choice of estimator with small excess risk is the *Empirical Risk Minimizer* (ERM). When $m = n/2$, if one chooses Q such that it outputs a Dirac around the ERM on a given sample, then under a BC with exponent β and for ‘parametric’ \mathcal{H} (such as the d -dimensional linear classifiers in Sec. 4), $\bar{L}(Q_{\leq m})$ and $\bar{L}(Q_{> m})$ are of order $\tilde{O}(n^{-1/(2-\beta)})$ w.h.p. [3, 16]. However, setting $P_n \equiv \delta(\text{ERM}(Z_{\leq n}))$ is not allowed, since otherwise $\text{KL}(P_n \| P_0) = \infty$. Instead one can choose P_n to be the generalized-Bayes/Gibbs posterior. In this case too, under a BC with exponent β and for parametric \mathcal{H} , the excess risk is of order $\tilde{O}(n^{-1/(2-\beta)})$ w.h.p. for clever choices of prior P_0 [3, 16].

6 Detailed Analysis

We start this section by presenting the convenient ESI notation and use it to present our main technical Lemma 10 (proofs of the ESI results are in Appendix B). We then continue with a proof of Theorem 1.

Definition 5. [ESI (Exponential Stochastic Inequality, pronounce as:easy) 16, 18] Let $\eta > 0$, and X, Y be any two random variables with joint distribution \mathbf{D} . We define

$$X \triangleq_{\eta}^{\mathbf{D}} Y \iff X - Y \triangleq_{\eta}^{\mathbf{D}} 0 \iff \mathbb{E}_{(X,Y) \sim \mathbf{D}} [e^{\eta(X-Y)}] \leq 1. \quad (8)$$

Definition 5 can be extended to the case where $\eta = \hat{\eta}$ is also a random variable, in which case the expectation in (8) needs to be replaced by the expectation over the joint distribution of $(X, Y, \hat{\eta})$. When no ambiguity can arise, we omit \mathbf{D} from the ESI notation. Besides simplifying notation, ESIs are useful in that they simultaneously capture “with high probability” and “in expectation” results:

Proposition 6. [ESI Implications] For fixed $\eta > 0$, if $X \triangleq_{\eta} Y$ then $\mathbb{E}[X] \leq \mathbb{E}[Y]$. For both fixed and random $\hat{\eta}$, if $X \triangleq_{\hat{\eta}} Y$, then $\forall \delta \in]0, 1[$, $X \leq Y + \frac{\ln \frac{1}{\delta}}{\hat{\eta}}$, with probability at least $1 - \delta$.

In the next proposition, we present two results concerning transitivity and additive properties of ESI:

Proposition 7. [ESI Transitivity and Chain Rule] (a) Let Z_1, \dots, Z_n be any random variables on \mathcal{Z}^n (not necessarily independent). If for some $(\gamma_i)_{i \in [n]} \in]0, +\infty[^n$, $Z_i \triangleq_{\gamma_i} 0$, for all $i \in [n]$, then

$$\sum_{i=1}^n Z_i \triangleq_{\rho_n} 0, \quad \text{where } \rho_n := \left(\sum_{i=1}^n \frac{1}{\gamma_i} \right)^{-1} \quad (\text{so if } \forall i \in [n], \gamma_i = \gamma > 0 \text{ then } \rho_n = \gamma/n). \quad (9)$$

(b) Suppose now that Z_1, \dots, Z_n are i.i.d. and let $X : \mathcal{Z} \times \bigcup_{i=1}^n \mathcal{Z}_i \rightarrow \mathbb{R}$ be any real-valued function. If for some $\eta > 0$, $X(Z_i; z_{<i}) \triangleq_{\eta} 0$, for all $i \in [n]$ and all $z_{<i} \in \mathcal{Z}^{i-1}$, then $\sum_{i=1}^n X(Z_i; Z_{<i}) \triangleq_{\eta} 0$.

We now give a basic PAC-Bayesian result for the ESI context (the proof, slightly different from standard change-of-measure arguments, is in Appendix B):

Proposition 8. [ESI PAC-Bayes] Fix $\eta > 0$ and $\{Y_h : h \in \mathcal{H}\}$ be any family of random variables such that for all $h \in \mathcal{H}$, $Y_h \triangleq_{\eta} 0$. Let P_0 be any distribution on \mathcal{H} and let $P : \bigcup_{i=1}^n \mathcal{Z}_i \rightarrow \Delta(\mathcal{H})$ be a learning algorithm. We have:

$$\mathbb{E}_{h \sim P_n} [Y_h] \triangleq_{\eta} \frac{\text{KL}(P_n \| P_0)}{\eta}, \quad \text{where } P_n := P(Z_{\leq n}). \quad (10)$$

In many applications (especially for our main result) it is desirable to work with a random (i.e. data-dependent) η in the ESI inequalities: one can obtain tighter bounds by tuning η in ‘hindsight’.

Proposition 9. [ESI from fixed to random η] Let \mathcal{G} be a countable subset of $]0, +\infty[$ and let π be a prior distribution over \mathcal{G} . Given a countable collection $\{Y_{\eta} : \eta \in \mathcal{G}\}$ of random variables satisfying $Y_{\eta} \triangleq_{\eta} 0$, for all fixed $\eta \in \mathcal{G}$, we have, for arbitrary estimator $\hat{\eta}$ with support on \mathcal{G} ,

$$Y_{\hat{\eta}} \triangleq_{\hat{\eta}} \frac{-\ln \pi(\hat{\eta})}{\hat{\eta}}. \quad (11)$$

The following key lemma, which is of independent interest, is central to our main result:

Lemma 10. [Key result: un-expected Bernstein] Let $X \sim \mathbf{D}$ be a random variable bounded from above by $b > 0$ almost surely, and let $\vartheta(u) := (-\ln(1-u) - u)/u^2$. For all $0 < \eta < 1/b$, we have

$$\mathbb{E}[X] - X \preceq_{\eta}^{\mathbf{D}} c \cdot X^2, \quad \text{for all } c \geq \eta \cdot \vartheta(\eta b). \quad (12)$$

The result is tight: for every $c < \eta \cdot \vartheta(\eta b)$, there exists a distribution \mathbf{D} so that (12) does not hold.

Lemma 10 is reminiscent of the following slight variation of Bernstein's inequality [10]: let X be any random variable bounded from below by $-b$, and let $\kappa(x) := (e^x - x - 1)/x^2$. For all $\eta > 0$, we have

$$\mathbb{E}[X] - X \preceq_{\eta} c \cdot \mathbb{E}[X^2], \quad \text{for all } c \geq \eta \cdot \kappa(\eta b). \quad (13)$$

Note that the un-expected Bernstein inequality in Lemma 10 has the X^2 lifted out of the expectation. In Appendix E, we prove (13) and compare it to standard versions of Bernstein. We also compare (12) to the related but distinct empirical Bernstein inequality due to [23, Theorem 4]. The detailed proof of Lemma 10 with the tight constants is (as far as we know) significantly harder than Bernstein's, and is postponed to the appendix. But it is easy to give a proof for bounded X with suboptimal constants:

Proof Sketch of Lemma 10 for $X \in [-1, 1]$ with suboptimal constants. Let $\nu > 0$ be such that $\nu\kappa(\nu) \leq 1 - \nu\kappa(\nu)$, e.g., $0 < \nu \leq 3/4$. We apply the standard Bernstein inequality (13) twice, once with X itself and once with X^2 (in the role of X), giving us, for $0 < \nu \leq 3/4$,

$$\begin{aligned} \mathbb{E}[X] - X \preceq_{\nu} \nu\kappa(\nu)\mathbb{E}[X^2] &\leq (1 - \nu\kappa(\nu))\mathbb{E}[X^2], \\ \mathbb{E}[X^2] - X^2 \preceq_{\nu} \nu\kappa(\nu)\mathbb{E}[X^4] &\leq \nu\kappa(\nu)\mathbb{E}[X^2], \quad \text{and so } (1 - \nu\kappa(\nu))\mathbb{E}[X^2] \preceq_{\nu} X^2. \end{aligned}$$

Chaining these ESIs using Prop. 7-(a) gives $\mathbb{E}[X] - X \preceq_{\nu/2} X^2$, so (12) holds with $c = 1$ and $\eta \leq 3/8$.

Proof of Theorem 1. Let $\eta \in]0, 1/b[$ and $c_{\eta} := \eta \cdot \vartheta(\eta b)$. For $1 \leq i \leq m < j \leq n$, define

$$\begin{aligned} X_h(Z_i; z_{>i}) &:= \ell_h(Z_i) - \mathbb{E}_{h' \sim Q(z_{>i})}[\ell_{h'}(Z_i)], \quad \text{for } z_{>i} \in \mathcal{Z}^{n-i}, \\ \tilde{X}_h(Z_j; z_{<j}) &:= \ell_h(Z_j) - \mathbb{E}_{h' \sim Q(z_{<j})}[\ell_{h'}(Z_j)], \quad \text{for } z_{<j} \in \mathcal{Z}^{j-1}. \end{aligned}$$

Since ℓ is bounded from above by b , Lemma 10 implies that $\forall h \in \mathcal{H}$ and $1 \leq i \leq m < j \leq n$,

$$\begin{aligned} \forall z_{>i} \in \mathcal{Z}^{n-i}, \quad Y_h^{\eta}(Z_i; z_{>i}) &:= \mathbb{E}_{Z'_i \sim \mathbf{D}}[X_h(Z'_i; z_{>i})] - X_h(Z_i; z_{>i}) - c_{\eta} \cdot X_h(Z_i; z_{>i})^2 \preceq_{\eta}^{Z_i \sim \mathbf{D}} 0, \\ \forall z_{<j} \in \mathcal{Z}^{j-1}, \quad \tilde{Y}_h^{\eta}(Z_j; z_{<j}) &:= \mathbb{E}_{Z'_j \sim \mathbf{D}}[\tilde{X}_h(Z'_j; z_{<j})] - \tilde{X}_h(Z_j; z_{<j}) - c_{\eta} \cdot \tilde{X}_h(Z_j; z_{<j})^2 \preceq_{\eta}^{Z_j \sim \mathbf{D}} 0, \end{aligned}$$

Since Z_1, \dots, Z_n are i.i.d. we can chain the ESIs above using Proposition 7-(b) to get:

$$S_1 := \sum_{i=1}^m Y_h^{\eta}(Z_i; Z_{>i}) \preceq_{\eta} 0, \quad S_2 := \sum_{j=m+1}^n \tilde{Y}_h^{\eta}(Z_j; Z_{<j}) \preceq_{\eta} 0. \quad (14)$$

Applying PAC-Bayes (Proposition 8) to S_1 and S_2 in (14) with priors $P(Z_{>m})$ and $P(Z_{\leq m})$, respectively, and common posterior $P_n = P(Z_{\leq n})$ on \mathcal{H} , we get, with $\text{KL}_{>m} := \text{KL}(P_n \| P(Z_{>m}))$ and $\text{KL}_{\leq m} := \text{KL}(P_n \| P(Z_{\leq m}))$:

$$\mathbb{E}_{h \sim P_n} \left[\sum_{i=1}^m Y_h^{\eta}(Z_i; Z_{>i}) \right] - \frac{\text{KL}_{>m}}{\eta} \preceq_{\eta} 0, \quad \mathbb{E}_{h \sim P_n} \left[\sum_{j=m+1}^n \tilde{Y}_h^{\eta}(Z_j; Z_{<j}) \right] - \frac{\text{KL}_{\leq m}}{\eta} \preceq_{\eta} 0.$$

We now apply Proposition 7-(a) to chain these two ESIs; this yields

$$\mathbb{E}_{h \sim P_n} \left[\sum_{i=1}^m Y_h^{\eta}(Z_i; Z_{>i}) + \sum_{j=m+1}^n \tilde{Y}_h^{\eta}(Z_j; Z_{<j}) \right] \preceq_{\frac{\hat{\eta}}{2}} \frac{\text{KL}(P_n \| P(Z_{>m})) + \text{KL}(P_n \| P(Z_{\leq m}))}{\eta}. \quad (15)$$

With the discrete prior π on \mathcal{G} , we have for any $\hat{\eta} = \hat{\eta}(Z_{\leq n}) \in \mathcal{G} \subset 1/b \cdot [1/\sqrt{n}, 1[$ (see Proposition 9),

$$\begin{aligned} \mathbb{E}_{h \sim P_n} \left[\sum_{i=1}^m Y_h^{\hat{\eta}}(Z_i; Z_{>i}) + \sum_{j=m+1}^n \tilde{Y}_h^{\hat{\eta}}(Z_j; Z_{<j}) \right] &\preceq_{\frac{\hat{\eta}}{2}} \frac{\text{COMP}_n}{\hat{\eta}} - \frac{2 \ln \pi(\hat{\eta})}{\hat{\eta}}, \quad \text{i.e.,} \\ n \cdot (L(P_n) - L_n(P_n)) &\preceq_{\frac{\hat{\eta}}{2}} n \cdot c_{\hat{\eta}} \cdot V_n + \frac{\text{COMP}_n + 2 \ln \frac{1}{\pi(\hat{\eta})}}{\hat{\eta}} + \\ &\left[\sum_{i=1}^m (\mathbb{E}_{Z'_i \sim \mathbf{D}}[\bar{\ell}_{Q_{>i}}(Z'_i)] - \bar{\ell}_{Q_{>i}}(Z_i)) + \sum_{j=m+1}^n (\mathbb{E}_{Z'_j \sim \mathbf{D}}[\bar{\ell}_{Q_{<j}}(Z'_j)] - \bar{\ell}_{Q_{<j}}(Z_j)) \right], \quad (16) \end{aligned}$$

where $\bar{\ell}_{Q_{>i}}(Z_i) := \mathbb{E}_{h' \sim Q(Z_{>i})}[\ell_{h'}(Z_i)]$ and $\bar{\ell}_{Q_{<j}}(Z_j) := \mathbb{E}_{h' \sim Q(Z_{<j})}[\ell_{h'}(Z_j)]$. Let U_n denote the quantity between the square brackets in (16). Using the un-expected Bernstein inequality (Lemma 10), together with Proposition 11, we get for any estimator $\hat{\nu}$ on \mathcal{G} :

$$U_n \leq_{\hat{\nu}} c_{\hat{\nu}} \cdot \left(\sum_{i=1}^m \mathbb{E}_{h' \sim Q(Z_{>i})}[\ell_{h'}(Z_i)^2] + \sum_{j=m+1}^n \mathbb{E}_{h' \sim Q(Z_{<j})}[\ell_{h'}(Z_j)^2] \right) + \frac{\ln \frac{1}{\pi(\hat{\nu})}}{\hat{\nu}}. \quad (17)$$

By chaining (17) and (16) using Proposition 7-(a) and dividing by n , we get:

$$L(P_n) \leq_{\frac{n\hat{\eta}\hat{\nu}}{\hat{\eta}+2\hat{\nu}}} L_n(P_n) + c_{\hat{\eta}} \cdot V_n + \frac{\text{COMP}_n + 2 \ln \frac{1}{\pi(\hat{\eta})}}{\hat{\eta} \cdot n} + c_{\hat{\nu}} \cdot V'_n + \frac{\ln \frac{1}{\pi(\hat{\nu})}}{\hat{\nu} \cdot n}. \quad (18)$$

We now apply Proposition 6 to (18) to obtain the following inequality with probability at least $1 - \delta$:

$$L(P_n) \leq L_n(P_n) + \left[c_{\hat{\eta}} \cdot V_n + \frac{\text{COMP}_n + 2 \ln \frac{1}{\pi(\hat{\eta}) \cdot \delta}}{\hat{\eta} \cdot n} \right] + \left\{ c_{\hat{\nu}} \cdot V'_n + \frac{\ln \frac{1}{\pi(\hat{\nu}) \cdot \delta}}{\hat{\nu} \cdot n} \right\}. \quad (19)$$

Inequality (3) follows after picking $\hat{\nu}$ and $\hat{\eta}$ to be, respectively, estimators which achieve the infimum over the closer of \mathcal{G} of the quantities between braces and square brackets in (19). \square

7 Conclusion and Future Work

The main goal of this paper was to introduce a new PAC-Bayesian bound based on a new proof technique; we also theoretically motivated the bound in terms of a Bernstein condition. The simple experiments we provided are to be considered as a basic sanity check — in future work, we plan to put the bound to real practical use by applying it to deep nets in the style of, *e.g.*, [36].

References

- [1] Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.
- [2] Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in neural information processing systems*, pages 9–16, 2007.
- [3] Jean-Yves Audibert. PAC-Bayesian statistical learning theory. *These de doctorat de l'Université Paris*, 6:29, 2004.
- [4] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- [5] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.
- [6] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.
- [7] Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 2003.
- [8] Olivier Catoni. *PAC-Bayesian Supervised Classification*. Lecture Notes-Monograph Series. IMS, 2007.
- [9] Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Lecture Notes-Monograph Series. IMS, 2007.
- [10] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, UK, 2006.
- [11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
- [12] Gintare K. Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*, 2017.

- [13] Tim Van Erven, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.
- [14] 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*, pages 353–360. ACM, 2009.
- [15] Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-François Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *The Journal of Machine Learning Research*, 16(1):787–860, 2015.
- [16] Peter D. Grünwald and Nishant A. Mehta. Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *Journal of Machine Learning Research*, 2019. accepted pending minor revision. Available as arXiv preprint arXiv:1605.00252.
- [17] Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.
- [18] Wouter M. Koolen, Peter D. Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems*, pages 4457–4465, 2016.
- [19] John Langford and Rich Caruana. (Not) bounding the true error. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 809–816. MIT Press, 2002.
- [20] John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.
- [21] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.
- [22] Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.
- [23] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings COLT 2009*, 2009.
- [24] David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh ACM Conference on Computational Learning Theory (COLT’ 98)*, pages 230–234. ACM Press, 1998.
- [25] David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT’ 99)*, pages 164–171. ACM Press, 1999.
- [26] David A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.
- [27] Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. In *Artificial Intelligence and Statistics*, pages 1085–1093, 2017.
- [28] Nishant A. Mehta and Robert C. Williamson. From stochastic mixability to fast rates. In *Advances in Neural Information Processing Systems*, pages 1197–1205, 2014.
- [29] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.
- [30] Behnam Neyshabur, Srinadh Bhojanapalli, David A. McAllester, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *ICLR*, 2018.
- [31] Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

- [32] Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.
- [33] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.
- [34] Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems*, pages 109–117, 2013.
- [35] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.
- [36] 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 *ICLR*, 2019.

A Additional Experiments

In this appendix, we run exactly the same experiments as in Section 4 of the main body, except that for the bounds we compare against, we build a prior from the first half of the data (*i.e.* we replace P_0 by $P(Z_{\leq m})$, where $m = n/2$) and use it to compute the bounds on the second half of the data. In this case, the ‘posterior’ distribution is $P(Z_{> m})$, and thus the term $\text{KL}(P_n \| P_0)$ is replaced by $\text{KL}(P(Z_{> m}) \| P(Z_{\leq m}))$. Recall that $P(Z_{> m}) \equiv \mathcal{N}(\hat{h}(Z_{> m}), \sigma^2 I_d)$, $P(Z_{\leq m}) \equiv \mathcal{N}(\hat{h}(Z_{\leq m}), \sigma^2 I_d)$, and $P(Z_{\leq n}) \equiv \mathcal{N}(\hat{h}(Z_{\leq n}), \sigma^2 I_d)$, where the variance σ^2 is learned from a geometric grid (see Section 4); our own bound is not affected by any of these changes. The results for the synthetic and UCI datasets are reported in Figure 2 and Table 2, respectively.

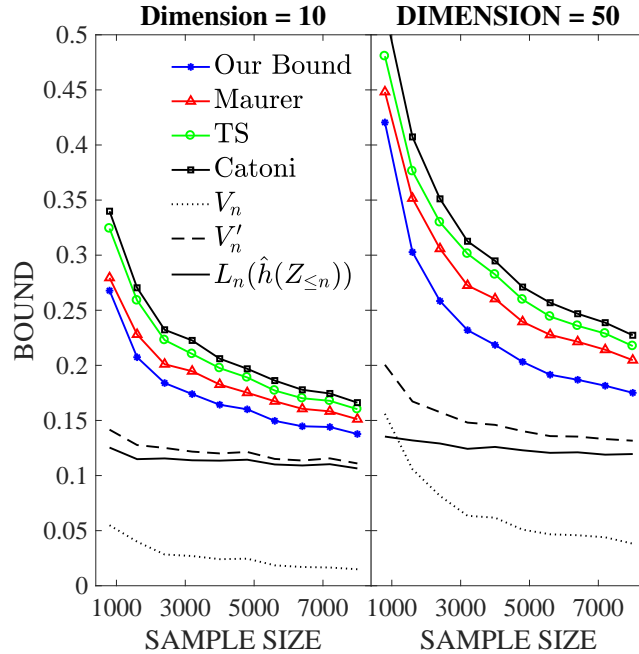


Figure 2: Results for the synthetic data.

Dataset	n	d	Test error of \hat{h}	Our	Maurer	TS	Catoni
Haberman	244	3	0.272	0.519	0.461	0.512	0.565
Breast-C.	560	9	0.068	0.185	0.177	0.236	0.239
Tic-Tac-Toe	766	27	0.046	0.190	0.225	0.276	0.291
Bank-note	1098	4	0.058	0.124	0.123	0.148	0.159
kr-vs-kp	2556	73	0.044	0.107	0.138	0.169	0.175
Spam-base	3680	57	0.173	0.297	0.341	0.380	0.389
Mushroom	6500	116	0.002	0.018	0.023	0.032	0.034
Adult	24130	108	0.168	0.195	0.214	0.220	0.226

Table 2: Results for the UCI datasets.

B Proofs for Section 6

Proof of Proposition 6. Let $Z = X - Y$. For fixed η , Jensen’s inequality yields $\mathbb{E}[Z] \leq 0$. For $\eta = \hat{\eta}$ that is either fixed or itself a random variable, applying Markov’s inequality to the random variable $e^{-\hat{\eta}Z}$ yields $Z \leq \frac{\ln \frac{1}{\delta}}{\hat{\eta}}$, with probability at least $1 - \delta$, for any $\delta \in]0, 1[$. \square

Proof of Proposition 7. [Part (a)] Fix $(\gamma_i)_{i \in [n]} \in]0, +\infty[^n$, and let $\rho_j := \left(\sum_{i=1}^j \frac{1}{\gamma_i}\right)^{-1}$, for $j \in [n]$. We proceed by induction to show that $\forall j \in [n]$, $\sum_{i=1}^j Z_i \preceq_{\rho_j} 0$. The result holds trivially for $j = 1$,

since $\rho_1 = \gamma_1$. Suppose that

$$\sum_{i=1}^j Z_i \trianglelefdeq_{\rho_j} 0, \quad (20)$$

for some $1 \leq j < n$. We now show that (20) holds for $j+1$; we have,

$$\begin{aligned} \mathbb{E} \left[e^{\frac{\rho_j \gamma_{j+1}}{\rho_j + \gamma_{j+1}} (\sum_{i=1}^j Z_i + Z_{j+1})} \right] &= \mathbb{E} \left[e^{\frac{\rho_j \gamma_{j+1}}{\rho_j + \gamma_{j+1}} \sum_{i=1}^j Z_i + \frac{\rho_j \gamma_{j+1}}{\rho_j + \gamma_{j+1}} Z_{j+1}} \right], \\ &\stackrel{\text{Jensen}}{\leq} \frac{\gamma_{j+1}}{\rho_j + \gamma_{j+1}} \mathbb{E} \left[e^{\rho_j \sum_{i=1}^j Z_i} \right] + \frac{\rho_j}{\rho_j + \gamma_{j+1}} \mathbb{E} \left[e^{\gamma_{j+1} Z_{j+1}} \right], \\ &\stackrel{\text{using (20)}}{\leq} 1. \end{aligned}$$

Thus the result holds for $j+1$, since $\rho_{j+1} = \frac{\rho_j \gamma_{j+1}}{\rho_j + \gamma_{j+1}}$. This establishes (9).

[Part (b)] This is a special case of [18, Lemma 6], who treat the general case with non-i.i.d. distributions. \square

Proof of Proposition 8. Let $\rho(h) = (dP_n/dP_0)(h)$ be the density of $h \in \mathcal{H}$ relative to the prior measure P_0 . We then have $\text{KL}(P_n \| P_0) = \mathbb{E}_{h \sim P_n} [\ln \rho(h)]$. We can now write:

$$\begin{aligned} \mathbb{E} \left[e^{\eta \mathbb{E}_{h \sim P_n} [Y_h] - \text{KL}(P_n \| P_0)} \right] &= \mathbb{E} \left[e^{\eta \mathbb{E}_{h \sim P_n} [Y_h - \log \rho(h)]} \right], \\ &\leq \mathbb{E} \left[\mathbb{E}_{h \sim P_n} \left[e^{\eta (Y_h - \log \rho(h))} \right] \right], \quad (\text{Jensen's Inequality}) \\ &= \mathbb{E} \left[\mathbb{E}_{h \sim P_n} \left[\frac{dP_0}{dP_n} \cdot e^{\eta Y_h} \right] \right], \\ &= \mathbb{E} \left[\mathbb{E}_{h \sim P_0} \left[e^{\eta Y_h} \right] \right], \\ &= \mathbb{E}_{h \sim P_0} \left[\mathbb{E} \left[e^{\eta Y_h} \right] \right], \quad (\text{Tonelli's Theorem}) \\ &= 1, \end{aligned}$$

where the final equality follows from our assumption that $Y_h \trianglelefdeq_{\eta} 0$, for all $h \in \mathcal{H}$. \square

Proof of Proposition 9. Since $Y_{\eta} \trianglelefdeq_{\eta} 0$, for $\eta \in \mathcal{G}$, we have in particular:

$$1 \geq \mathbb{E} \left[\sum_{\eta \in \mathcal{G}} \pi(\eta) e^{\eta Y_{\eta}} \right] \geq \mathbb{E} \left[\pi(\hat{\eta}) e^{\hat{\eta} Y_{\hat{\eta}}} \right], \quad (21)$$

where the right-most inequality follows from the fact that the expectation of a countable sum of positive random variable is greater than the expectation of a single element in the sum. Rearranging (21) gives (11). \square

C Proof of Theorem 4

In what follows, for $h \in \mathcal{H}$, we denote $X_h(Z) := \ell_h(Z) - \ell_{h_*}(Z)$ the excess loss random variable, where h_* is the risk minimizer within \mathcal{H} . Let

$$\epsilon(\eta) := \frac{1}{\eta} \ln \mathbb{E}_{Z \sim \mathbf{D}} \left[e^{-\eta X_h(Z)} \right]$$

be its *normalized cumulant generating function*. We need the following useful lemmas:

Lemma 11. [18] *Let $h \in \mathcal{H}$ and X_h be as above. Then, for all $\eta \geq 0$,*

$$s_{\eta} \cdot X_h(Z)^2 - X_h(Z) \trianglelefdeq_{\eta} \epsilon(2\eta) + s_{\eta} \cdot \epsilon(2\eta)^2, \quad \text{where } s_{\eta} := \frac{\eta}{1 + \sqrt{1 + 4\eta^2}}.$$

Lemma 12. [18] *Let $b > 0$, and suppose that $X_h \in [-b, b]$ almost surely, for all $h \in \mathcal{H}$. If the (β, B) -Bernstein condition holds with $\beta \in [0, 1]$ and $B > 0$, then*

$$\epsilon(\eta) \leq (B\eta)^{\frac{1}{1-\beta}}, \quad \text{for all } \eta \in]0, 1/b].$$

Lemma 13. [10] Let $b > 0$, and suppose that $X_h \in [-b, b]$ almost surely, for all $h \in \mathcal{H}$. Then

$$\epsilon(\eta) \leq \frac{\eta b^2}{2}, \quad \text{for all } \eta \in \mathbb{R}.$$

Proof of Theorem 4. First we apply the following inequality

$$(a - d)^2 \leq 2(a - c)^2 + 2(d - c)^2 \quad (22)$$

which holds for all $a, c, d \in \mathbb{R}$ to upper bound V_n . Let's focus on the first term in the expression of V_n , which we denote V_n^{left} : that is,

$$V_n^{\text{left}} := \mathbb{E}_{h \sim P_n} \left[\frac{1}{n} \sum_{i=1}^m (\ell_h(Z_i) - \mathbb{E}_{h' \sim Q(Z_{>i})} [\ell_{h'}(Z_i)])^2 \right]. \quad (23)$$

Letting $X_h(Z) := \ell_h(Z) - \ell_{h_*}(Z)$ and applying (22) with $a = \ell_h(Z_i)$, $c = \ell_{h_*}(Z_i)$, and $d = \mathbb{E}_{h' \sim Q(Z_{>i})} [\ell_{h'}(Z_i)] \stackrel{*}{=} \mathbb{E}_{h' \sim Q(Z_{>m})} [\ell_{h'}(Z_i)]$ (where $\stackrel{*}{=}$ is due to our assumption on Q), we get:

$$\begin{aligned} V_n^{\text{left}} &\leq \mathbb{E}_{h \sim P_n} \left[\frac{2}{n} \sum_{i=1}^m X_h(Z_i)^2 \right] + \frac{2}{n} \sum_{i=1}^m (\mathbb{E}_{h' \sim Q(Z_{>m})} [\ell_{h'}(Z_i)] - \ell_{h_*}(Z_i))^2, \\ &\leq \mathbb{E}_{h \sim P_n} \left[\frac{2}{n} \sum_{i=1}^m X_h(Z_i)^2 \right] + \mathbb{E}_{h \sim Q(Z_{>m})} \left[\frac{2}{n} \sum_{i=1}^m X_h(Z_i)^2 \right]. \quad (\text{by Jensen's Inequality}) \end{aligned} \quad (24)$$

Let $i \in [m]$, $h \in \mathcal{H}$, and $\eta \in]0, 1/b[$. Under the (β, B) -Bernstein condition, Lemmas 11-13 imply,

$$s_\eta \cdot X_h(Z_i)^2 \triangleq_\eta X_h(Z_i) + \left(1 + \frac{b}{2}\right) (2B\eta)^{\frac{1}{1-\beta}}, \quad (25)$$

where $s_\eta := \eta / (1 + \sqrt{1 + 4\eta^2})$. Now, due to the Bernstein inequality (13), we have

$$\begin{aligned} X_h(Z_i) &\triangleq_\eta \mathbb{E}_{Z'_i \sim \mathbf{D}} [X_h(Z'_i)] + \tilde{c}_\eta \cdot \mathbb{E}_{Z'_i \sim \mathbf{D}} [X_h(Z'_i)^2], \quad \text{where } \tilde{c}_\eta := \eta \cdot \kappa(\eta b), \\ &\triangleq \mathbb{E}_{Z'_i \sim \mathbf{D}} [X_h(Z'_i)] + \tilde{c}_\eta \cdot \mathbb{E}_{Z'_i \sim \mathbf{D}} [X_h(Z'_i)]^\beta, \quad (\text{by the Bernstein condition}) \\ &\triangleq_\eta 2\mathbb{E}_{Z'_i \sim \mathbf{D}} [X_h(Z'_i)] + a_\beta^{\frac{\beta}{1-\beta}} \cdot (\tilde{c}_\eta)^{\frac{1}{1-\beta}}, \quad \text{where } a_\beta := (1 - \beta)^{1-\beta} \beta^\beta. \end{aligned} \quad (26)$$

The last inequality follows by the fact that $z^\beta = a_\beta \cdot \inf_{\nu > 0} z/\nu + \nu^{\frac{\beta}{1-\beta}}$, for $z \geq 0$ (in our case, we set $\nu = a_\beta \cdot \tilde{c}_\eta$ to get to (26)). By chaining (25) with (26) using Proposition 7-(a), we get:

$$\begin{aligned} s_\eta \cdot X_h(Z_i)^2 &\triangleq_{\frac{\eta}{2}} 2\mathbb{E}_{Z'_i \sim \mathbf{D}} [X_h(Z'_i)] + a_\beta^{\frac{\beta}{1-\beta}} \cdot (\tilde{c}_\eta)^{\frac{1}{1-\beta}} + \left(1 + \frac{b}{2}\right) (2B\eta)^{\frac{1}{1-\beta}}. \\ &\triangleq_{\frac{\eta}{2}} 2\mathbb{E}_{Z'_i \sim \mathbf{D}} [X_h(Z'_i)] + \mathcal{P} \cdot \eta^{\frac{1}{1-\beta}}, \quad \text{where } \mathcal{P} := a_\beta^{\frac{\beta}{1-\beta}} + \left(1 + \frac{b}{2}\right) (2B)^{\frac{1}{1-\beta}}, \end{aligned} \quad (27)$$

where in the last inequality we used $\kappa(1) \leq 1$. Since (27) holds for all $h \in \mathcal{H}$, it still holds in expectation over \mathcal{H} with respect to the distribution $Q(Z_{>m})$ (recall that $i \leq m$);

$$s_\eta \cdot \mathbb{E}_{h \sim Q(Z_{>m})} [X_h(Z_i)^2] \triangleq_{\frac{\eta}{2}} 2\mathbb{E}_{h \sim Q(Z_{>m})} [\mathbb{E}_{Z'_i \sim \mathbf{D}} [X_h(Z'_i)]] + \mathcal{P} \cdot \eta^{\frac{1}{1-\beta}}. \quad (28)$$

Since the samples $Z_{\leq n}$ are i.i.d, we have $\mathbb{E}_{Z_i \sim \mathbf{D}} [\ell_h(Z_i)] = \mathbb{E}_{Z_j \sim \mathbf{D}} [\ell_h(Z_j)]$, for all $i, j \in [m]$. Thus, after summing (27) and (28), for $i = 1, \dots, m$, using Proposition 7-(b) and dividing by n , we get

$$\frac{s_\eta}{n} \sum_{i=1}^m X_h(Z_i)^2 \triangleq_{\frac{n-\eta}{2}} \mathbb{E}_{Z \sim \mathbf{D}} [X_h(Z)] + \frac{\mathcal{P}}{2} \cdot \eta^{\frac{1}{1-\beta}}, \quad (29)$$

$$\mathbb{E}_{h \sim Q(Z_{>m})} \left[\frac{s_\eta}{n} \sum_{i=1}^m X_h(Z_i)^2 \right] \triangleq_{\frac{n-\eta}{2}} \mathbb{E}_{h \sim Q(Z_{>m})} [\mathbb{E}_{Z \sim \mathbf{D}} [X_h(Z)]] + \frac{\mathcal{P}}{2} \cdot \eta^{\frac{1}{1-\beta}}. \quad (m = n/2) \quad (30)$$

Now we apply PAC-Bayes (Proposition 8) to (29), with prior $P(Z_{>m})$ and posterior P_n , and obtain:

$$\mathbb{E}_{h \sim P_n} \left[\frac{s_\eta}{n} \sum_{i=1}^m X_h(Z_i)^2 \right] \triangleq_{\frac{n-\eta}{2}} \mathbb{E}_{h \sim P_n} [\mathbb{E}_{Z \sim \mathbf{D}} [X_h(Z)]] + \frac{\mathcal{P}}{2} \cdot \eta^{\frac{1}{1-\beta}} + \frac{2\text{KL}(P_n \| P(Z_{>m}))}{\eta \cdot n}. \quad (31)$$

Note that the upper-bound on V_n^{left} in (24) is the sum of left-hand sides of (30) and (31) divided by $s_\eta/2$. From now on, we restrict η to the range $]0, 1/(2b)[$ and define

$$\mathcal{A}_\eta := \frac{2c_\eta}{s_\eta} \leq 2\vartheta \left(\frac{1}{2}\right) \cdot \left(1 + \sqrt{1 + \frac{1}{b^2}}\right) =: \mathcal{A}, \quad \text{for } \eta \in]0, \frac{1}{2b}[.$$

Chaining (30) and (31) using Proposition 7-(a) and multiplying throughout by \mathcal{A}_η , yields

$$c_\eta \cdot V_n^{\text{left}} \leq \frac{n\eta}{4\mathcal{A}_\eta} \mathcal{A} \cdot (\bar{L}(P_n) + \bar{L}(Q(Z_{>m}))) + \mathcal{P}\mathcal{A}\eta^{\frac{1}{1-\beta}} + \frac{2\mathcal{A} \cdot \text{KL}(P_n \| P(Z_{>m}))}{\eta \cdot n}. \quad (32)$$

By a symmetric argument, a version of (32), with $Q(Z_{>m})$ [resp. $P(Z_{>m})$] replaced by $Q(Z_{\leq m})$ [resp. $P(Z_{\leq m})$], holds for $V_n^{\text{right}} := V_n - V_n^{\text{left}}$. Using Proposition 7-(a) again, to chain the ESI inequalities of $c_\eta \cdot V_n^{\text{left}}$ and $c_\eta \cdot V_n^{\text{right}}$, we obtain:

$$c_\eta \cdot V_n \leq \frac{n\eta}{8\mathcal{A}_\eta} \mathcal{A} \cdot (2\bar{L}(P_n) + \bar{L}(Q_{\leq m}) + \bar{L}(Q_{>m})) + 2\mathcal{P}\mathcal{A}\eta^{\frac{1}{1-\beta}} + \frac{2\mathcal{A} \cdot \text{COMP}_n}{\eta \cdot n}, \quad (33)$$

where $Q_{>m} := Q(Z_{>m})$ and $Q_{\leq m} := Q(Z_{\leq m})$. Let $\delta \in]0, 1[$, and π and \mathcal{G} be as in (4). Applying Proposition 9 to (33) to obtain the corresponding ESI inequality with a random estimator $\hat{\eta} = \hat{\eta}(Z_{\leq n})$ with support on \mathcal{G} , and then applying Proposition 6, we get, with probability at least $1 - \delta$,

$$c_{\hat{\eta}} \cdot V_n \leq \mathcal{A} \cdot (2\bar{L}(P_n) + \bar{L}(Q_{\leq m}) + \bar{L}(Q_{>m})) + 2\mathcal{P}\mathcal{A}\hat{\eta}^{\frac{1}{1-\beta}} + \frac{2\mathcal{A} \cdot \text{COMP}_n + 8\mathcal{A} \ln \frac{|\mathcal{G}|}{\delta}}{\hat{\eta} \cdot n}. \quad (34)$$

Now adding $(\text{COMP}_n + s_{\delta,n})/(\hat{\eta} \cdot n)$ on both sides of (34) and choosing the estimator $\hat{\eta}$ optimally in the closure of \mathcal{G} yields the desired result. \square

D Proof of Lemma 10

Proof. Although the following proof does not use any of the results in [28] directly, the basic arguments are heavily inspired by the developments in that paper. We will show a slight extension of Lemma 10, namely that, (a) for all $0 < u < 1$:

$$\sup_{\rho \leq u} \sup_{P: \mathbb{E}_P[X] = \rho, P(X \leq u) = 1} \mathbb{E}_{X \sim P} \left[e^{E[X] - X - cX^2} \right] \leq 1 \text{ if } c \geq \vartheta(u), \quad (35)$$

and (b), for all $\beta > 0, u > 0$,

$$\sup_{\rho \leq u} \sup_{P: \mathbb{E}_P[X] = \rho, P(X \leq u) = 1} \mathbb{E}_{X \sim P} \left[e^{\beta E[X] - X - cX^2} \right] > 1 \text{ if } 0 < c < \vartheta(u) \text{ or } \beta \neq 1. \quad (36)$$

The statement of the lemma (12) follows by replacing X in (35) by ηX and set u to $u := \eta b < 1$; tightness follows by (36) for $\beta = 1$ and the same choices of X and u .

Stage 1. We start by showing that

$$\sup_{\rho \leq u} \sup_{P: \mathbb{E}_P[X] = \rho, P(X \leq u) = 1} g_{c,\beta}(P), \quad \text{with } g_{c,\beta}(P) := \mathbb{E}_{X \sim P} \left[e^{\beta \rho - X - cX^2} \right] \quad (37)$$

is achieved for a distribution P with support on just two points $\underline{x} \leq \bar{x} \leq u$. To see this, first consider any distribution P satisfying the constraint $P(X \leq u) = 1$, and let $\rho := \mathbb{E}_P[X]$. Let \underline{u} be the unique point that satisfies $\underline{u} < 0$ and $\beta\rho - \underline{u} - c\underline{u}^2 = 0$, and consider the event $\mathcal{E} := \{X \leq \underline{u}\}$.

Let P' be the distribution that is identical to P on $X > \underline{u}$ but that has all mass on \mathcal{E} transferred to \underline{u} . That is, $P'(\mathcal{E}) = P'(X = \underline{u}) = P(\mathcal{E})$, $P'(X | \mathcal{E}^c) = P(X | \mathcal{E}^c)$, where \mathcal{E}^c denotes the complement of \mathcal{E} . Then P' still satisfies the constraint $P'(X \leq u) = 1$, yet $\mathbb{E}_{P'}[X] \geq \rho$ and $\mathbb{E}_{P'}[\exp(\rho - X - cX^2)] > \mathbb{E}_P[\exp(\rho - X - cX^2)]$ so $g_{c,\beta}(P') \geq g_{c,\beta}(P)$. It follows that without loss of generality, we can restrict the supremum in (37) to be on distributions such that all points with $X < 0$ also have $X \geq \underline{u}$. Thus, the supremum becomes over distributions with compact support $[\underline{u}, u]$.

Now consider integer $M > 0$ and suppose that, for given fixed ρ , we restrict the supremum over P to be on distributions with expectation ρ and support on an equally spaced grid of M points

$\underline{u} = u_1 < u_2 < u_3 < \dots < u_M = u$. Since we can assume compact support $[\underline{u}, u]$ and the objective $g_{c,\beta}$ is smooth and bounded, then for every $\epsilon > 0$, there exists a finite M_ϵ such that the P achieving the maximum within this restricted optimization problem is within ϵ of the supremum of $g_{c,\beta}(P)$ over all P satisfying the constraint $P(X \leq u) = 1$. But in the restricted maximization, both the constraint and the objective are linear in P (viewed as an M_ϵ -dimensional probability vector), given by two $(M_\epsilon - 1)$ -dimensional hyperplanes in M_ϵ -dimensional space. It follows that the maximum must be achieved at a P_ϵ that is an extreme point of the intersection of the simplex in \mathbb{R}^{M_ϵ} and the hyperplane corresponding the constraint $P(X \leq u) = 1$, which is a vector with at most two non-zero components. It follows that the supremum is in fact achieved for a P with support on just two points, say \underline{x} and \bar{x} , with $\underline{u} \leq \underline{x} \leq \bar{x} \leq u$.

Stage 2. We next show that, for any $\beta \neq 1$, $\sup g_{c,\beta}(P) > 1$, and that without loss of generality we can assume $\underline{x} \leq 0, \bar{x} \geq 0$. We can write $g_{c,\beta}(P)$ as

$$p \cdot e^{-\underline{x} + \beta \rho - c \underline{x}^2} + (1-p) e^{-\bar{x} + \beta \rho - c \bar{x}^2}$$

with $\rho = \mathbb{E}_P[X]$, and we need to maximize this over $\rho = p\underline{x} + (1-p)\bar{x}$, so that in the end, we want to maximize over $0 \leq p \leq 1, \underline{u} \leq \underline{x} \leq \bar{x} \leq u$, the expression

$$p \cdot e^{-\underline{x} + \beta(p\underline{x} + (1-p)\bar{x}) - c \underline{x}^2} + (1-p) e^{-\bar{x} + \beta(p\underline{x} + (1-p)\bar{x}) - c \bar{x}^2}$$

Now we write $\underline{x} = \bar{x} - a$ for some $a \geq 0$. The expression becomes

$$p \cdot e^{-\beta pa + (\beta-1)\bar{x} + a - c(\bar{x}-a)^2} + (1-p) \cdot e^{-\beta pa + (\beta-1)\bar{x} - c\bar{x}^2}$$

which is equal to

$$f(p, a, \bar{x}) := e^{-c\bar{x}^2 - \beta pa + (\beta-1)\bar{x}} \left(p e^{a+2ca\bar{x}-ca^2} + 1 - p \right) \stackrel{\text{if } \beta=1}{=} e^{c\bar{x}^2 - pa} \left(p e^{a+2ca\bar{x}-ca^2} + 1 - p \right), \quad (38)$$

where the dependency of f on c and β is suppressed in the notation. At $p = 1$ and $p = 0$, this simplifies to (using also \underline{x} again)

$$f(1, a, \bar{x}) = e^{-c\bar{x}^2 - \beta a + (\beta-1)\bar{x}} \left(e^{a+2ca\bar{x}-ca^2} \right) = e^{-c\bar{x}^2 + (\beta-1)\bar{x}} \stackrel{\text{if } \beta=1}{=} e^{-c\bar{x}^2} \quad (39)$$

$$f(0, a, \bar{x}) = e^{-c\bar{x}^2 + (\beta-1)\bar{x}} \stackrel{\text{if } \beta=1}{=} e^{-c\bar{x}^2}. \quad (40)$$

If $\beta < 1$, we can choose $\underline{x} = \bar{x} - a$ negative yet very close to 0 making $f(1, a, \bar{x}) > 1$; if $\beta > 1$, we can choose \bar{x} positive yet very close to 0 making $f(0, a, \bar{x}) > 1$. Thus, $\sup g_{c,\beta}(P)$ can be made larger than 1 by P satisfying the constraint if $\beta \neq 1$. This shows (36) for the case $\beta \neq 1$; it remains to study the supremum of $g_{c,\beta}(P)$ over all allowed P with $\beta = 1$, or, equivalently, the supremum of $f(p, a, \bar{x})$ for $\beta = 1$ over $0 \leq p \leq 1$ and \underline{x}, \bar{x} with $\underline{u} \leq \underline{x} \leq \bar{x} \leq u, a = \bar{x} - \underline{x}$. But with $\beta = 1$, without loss of generality, we can take $\bar{x} \geq 0$, since if $\bar{x} < 0$, then replacing \bar{x} by $-\bar{x}$ on the right of (38) makes the expression strictly larger. Similarly one shows that without loss of generality one can take $\underline{x} \leq 0$ by re-expressing $f(p, a, \bar{x})$ in terms of $\underline{x} = \bar{x} - a$. We may thus, from now on, assume $\beta = 1, \underline{x} \leq 0, \bar{x} \geq 0, a \geq 0$.

Stage 3. We now show that the supremum $f(p, a, \bar{x})$ over the allowed p, a, \bar{x} is necessarily larger than 1 if c is too small; more precisely, we show that (36) holds for the case $\beta = 1$. Given the previous, it suffices to determine the maximum over (38) for $a \geq \bar{x}$ and $0 \leq p \leq 1$, for each given $0 \leq \bar{x} \leq u$. The partial derivatives to p and a are:

$$\begin{aligned} \frac{\partial}{\partial p} f(p, a, \bar{x}) &= e^{-c\bar{x}^2 - pa} \left(\left(e^{a+2ca\bar{x}-ca^2} - 1 \right) - a \cdot \left(p e^{a+2ca\bar{x}-ca^2} + (1-p) \right) \right) \\ &= e^{-c\bar{x}^2 - pa} \left(e^{a+2ca\bar{x}-ca^2} (1-ap) - 1 - a + ap \right) \end{aligned} \quad (41)$$

$$\begin{aligned} \frac{\partial}{\partial a} f(p, a, \bar{x}) &= -p \cdot e^{-c\bar{x}^2 - pa} \left(p e^{a+2ca\bar{x}-ca^2} + (1-p) \right) + \\ &+ e^{-c\bar{x}^2 - pa} \cdot p \cdot e^{a+2ca\bar{x}-ca^2} \cdot (1+2c\bar{x}-2ca) \\ &= p(1-p) \cdot e^{-c\bar{x}^2 - pa} \cdot \left(-1 + e^{a+2ca\bar{x}-ca^2} \left(1 + 2c \frac{\bar{x}-a}{1-p} \right) \right). \end{aligned} \quad (42)$$

At $a = \bar{x}$ (i.e. $\underline{x} = 0$), $f(p, a, \bar{x})$ simplifies to

$$f(p, \bar{x}, \bar{x}) = e^{-c\bar{x}^2 - p\bar{x}} \cdot (pe^{\bar{x} + c\bar{x}^2} + (1 - p)) \text{ so } f(1, \bar{x}, \bar{x}) = 1$$

and the partial derivative to p at $(p, a, \bar{x}) = (1, \bar{x}, \bar{x})$ becomes

$$e^{-c\bar{x}^2 - \bar{x}} \left((e^{\bar{x} + c(\bar{x})^2} - 1) - \bar{x}e^{\bar{x} + c(\bar{x})^2} \right) = 1 - e^{-c\bar{x}^2 - \bar{x}} - \bar{x}. \quad (43)$$

If (43) is negative, we can take $a = \bar{x}$ and p slightly smaller than 1 to get $f(p, a, \bar{x}) > 1$. This happens if and only if c is smaller than

$$\frac{-\ln(1 - \bar{x}) - \bar{x}}{\bar{x}^2} = \vartheta(\bar{x}). \quad (44)$$

By taking $\bar{x} = a = u$, and p slightly smaller than 1 again, we get $f(p, a, \bar{x}) > 1$ if $c < \vartheta(u)$; this shows (36) for the case $\beta = 1$.

Stage 4. We will now show that if c is not smaller than (44), then

$$f(p, a, \bar{x}) \leq 1 \text{ for all } 0 \leq p \leq 1, 0 \leq \bar{x} < 1, a \geq \bar{x}; \quad (45)$$

from this (35), and hence the result, follows. (the constraint $\bar{x} < 1$ follows because we assume $\bar{x} \leq u < 1$). Since the supremum over $g_{c,\beta}(P)$ or equivalently $f(p, a, \bar{x})$ is decreasing in c , it is sufficient to show (45) for c equal to (44), i.e. for

$$c\bar{x}^2 = g(\bar{x}) := -\ln(1 - \bar{x}) - \bar{x} \quad (46)$$

While the required bound on f is very easy to show numerically by plotting graphs, a formal proof requires some work: using $\beta = 1$, we see from (39) and (40) that, if $f(p, a, \bar{x}) > 1$, this has to happen at some $0 < p < 1$. We already know that we can assume $a \geq 0$; at $a = 0$ we have $f(p, a, \bar{x}) \leq 1$ for all $0 < p < 1$ so if $f(p, a, \bar{x}) > 1$, we can furthermore assume that this happens at some $a > 0$, $0 < p < 1$. From (42) we see that, if $\bar{x} - a < -(1 - p)/2c$, then the derivative $(\partial/\partial a)f(p, a, \bar{x})$ is negative uniformly for $0 < p < 1$, so we have for every $\delta > 0$: if $f(p, a, \bar{x}) > 1$, this has to happen at some (p^*, a^*) with $0 < p^* < 1$ and $0 < a^* < (1 - p^*)/2c + \bar{x} + \delta$. By taking $\delta = 1 - \bar{x} > 0$, we find that $(p^*, a^*) \in (0, 1) \times (0, 1/2c + 1)$. Since at the boundaries of this open rectangle, we either have $f(p, a, \bar{x}) \leq 1$ or (at $a = 1/2c + 1$) the partial derivative $(\partial/\partial a)f(p, a, \bar{x}) < 0$, it follows that if there is a p^*, a^*, \bar{x} with $f(p^*, a^*, \bar{x}) > 1$ at all, there is also an (p, a) inside the open rectangle $(0, 1) \times (0, 1/2c + 1)$ with $f(p, a, \bar{x}) > 1$ such that *both partial derivatives of f to p and a are equal to 0 at (p, a)* . Using (41) and (42) we get for this (p, a, \bar{x}) :

$$\begin{aligned} \frac{1 + a - ap}{1 - ap} &= e^{a + 2ca\bar{x} - ca^2} \\ \frac{1}{1 + 2c\frac{\bar{x} - a}{1 - p}} &= e^{a + 2ca\bar{x} - ca^2}. \end{aligned} \quad (47)$$

Setting the left-hand sides equal and solving for $1 - p$ gives

$$\frac{2c(\bar{x} - a)}{1 - p} = -a(1 + 2c(\bar{x} - a)). \quad (48)$$

If (48) has no solution within the rectangle $(p, a, \bar{x}) \in (0, 1) \times (0, 1/2c + 1) \times [0, 1]$, then $f(p, a, \bar{x}) \leq 1$ at all allowed p, a and \bar{x} and we are done; so henceforth we assume that (48) (and (47)) do have a solution within the rectangle. We now show that from (46) and (48) we then also get the further constraints that

$$0 < \bar{x} \text{ and } a > \bar{x}, \text{ so } (p, a, \bar{x}) \in (0, 1) \times (\bar{x}, 1/2c + 1) \times (0, 1). \quad (49)$$

Here $\bar{x} > 0$ follows from (48) since $\bar{x} \geq 0$ and (from (46) and the fact that g is increasing) $c \geq 1/2$. Thus $\bar{x} = 0$ would imply $2c = (1 - p)(1 - 2ca)$, which, since $a > 0$, is impossible. We also know that $\underline{x} \leq 0$ hence $a \geq \bar{x} > 0$; but if $a = \bar{x}$, (48) leads to a contradiction, so $a > \bar{x}$ and (49) follows.

We thus assume that both (48) and (47) hold for some (p, a, \bar{x}) as in (49). Combining the two equalities gives

$$1 - a(1 + 2c(\bar{x} - a)) = e^{-a - 2ca\bar{x} + ca^2} = e^{-a(1 + 2c(\bar{x} - a)) - ca^2}.$$

We now want to eliminate the variable c , using (46), which gives a convenient expression for $c\bar{x}^2$ in terms of the function g defined there. To rewrite the above in terms of $c\bar{x}^2$, we perform a change of variables, writing $\alpha = a/\bar{x}$, which is possible since we already derived $a > \bar{x} > 0$; note that $\alpha > 1$. This gives

$$1 - \alpha\bar{x} \left(1 + \frac{2c\bar{x}^2(1-\alpha)}{\bar{x}} \right) = e^{-\alpha\bar{x} - c\bar{x}^2(2\alpha-\alpha^2)} \quad (50)$$

i.e.

$$1 - \alpha(\bar{x} + 2g(\bar{x})(1-\alpha)) = e^{-\alpha\bar{x} - g(\bar{x})(2\alpha-\alpha^2)} = e^{-\alpha(\bar{x} + 2g(\bar{x})(1-\alpha)) - \alpha^2 g(\bar{x})}, \quad (51)$$

We also note that, rewritten with our change of variables, after some rearranging, (48) gives $\frac{2c(\alpha-1)}{1-p} = \alpha(1 - 2c\bar{x}(\alpha-1))$, which, since $\alpha > 1, \bar{x} > 0, c \geq 1/2$ implies $1 - 2c\bar{x}(\alpha-1) > 0$, *i.e.*

$$\alpha - 1 < \frac{1}{2c\bar{x}} \leq 1/\bar{x} \text{ and, via (50), } e^{-\alpha\bar{x} - c\bar{x}^2(2\alpha-\alpha^2)} < 1. \quad (52)$$

It thus suffices to show that (51) cannot hold for $1 < \alpha < 1/\bar{x} + 1, 0 < \bar{x} < 1$ (for then there can be no (a, p, \bar{x}) satisfying (49) that also satisfies (50), *i.e.* that satisfies (47), *i.e.* that has $f(p, a, \bar{x}) > 1$, and the result would be proved). We show this by comparing the left-hand and right-hand side of (51) for any fixed α satisfying $\alpha > 1$ and (52). At $\bar{x} = 0$, both sides are equal and also their derivatives towards \bar{x} are equal. The second derivative towards \bar{x} of the left-hand side is given by

$$-2\alpha(1-\alpha)g''(\bar{x}) = -2\alpha(1-\alpha)\frac{1}{(1-\bar{x})^2}, \quad (53)$$

whereas the second derivative for the right-hand side is given by

$$\begin{aligned} & e^{-\alpha(\bar{x} + 2g(\bar{x})(1-\alpha)) - \alpha^2 g(\bar{x})} \left(\alpha^2(1 + 2g'(\bar{x})(1-\alpha) + \alpha g'(\bar{x}))^2 - \alpha^2 - 2\alpha(1-\alpha)g''(\bar{x}) - \alpha^2 g''(\bar{x}) \right) \\ &= e^{-\alpha(\bar{x} + 2g(\bar{x})(1-\alpha)) - \alpha^2 g(\bar{x})} \cdot \left(\frac{\alpha^2((1-x(\alpha-1))^2 - 1)}{(1-\bar{x})^2} - 2\alpha(1-\alpha)\frac{1}{(1-\bar{x})^2} \right). \end{aligned} \quad (54)$$

By our constraint on x and α , we have $0 < 1 - x(\alpha-1) < 1$, so that the expression between brackets is strictly smaller than the second derivative (53), which, since $\alpha > 1$, is positive. Since, by (52), the first (exponential) factor in (54) is upper bounded by 1, it follows that the second derivative of the left-hand side of (51) is strictly larger than the one on the right-hand side at all $\bar{x} > 0$; this shows that there can be no (p, a, \bar{x}) satisfying (49) that also satisfies (50), *i.e.* that satisfies (47), *i.e.* that has $f(p, a, \bar{x}) > 1$; the result is proved. \square

E Comparison Between ‘Bernstein’ Inequalities

Discussion and Proof of Our Version of Bernstein’s Inequality (13). Standard versions of Bernstein’s inequality (see [10], and [13, Lemma 5.6]) can also be brought in ESI notation. In particular, compared with our version they express the inequality in terms of the random variable $Y = -X$, which is then upper bounded by b ; more importantly, they have the second moment rather than the variance on the right-hand side, resulting in a slightly worse multiplicative factor $\kappa(2\eta b)$ instead of our $\kappa(\eta b)$; the proof is a standard one (see [10, Lemma A.4]) with trivial modifications: let $U := \eta X$ and $\bar{u} := \eta b$. Since $\kappa(u)$ is nondecreasing in u and $U \leq \bar{u}$, we have

$$\frac{e^U - U - 1}{U^2} \leq \frac{e^{\bar{u}} - \bar{u} - 1}{\bar{u}^2},$$

and hence $e^U - U - 1 \leq \kappa(\bar{u})U^2$. Taking expectation on both sides and using that $\ln \mathbb{E}[e^U] \leq \mathbb{E}[U] - 1$, we get $\ln \mathbb{E}[e^U] - \mathbb{E}[U] \leq \kappa(\bar{u})\mathbb{E}[U^2]$. The result follows by exponentiating, rearranging, and using the ESI definition.

Comparison Between Un-expected and Empirical Bernstein Inequalities. The proof of the following proposition demonstrates how the un-expected Bernstein inequality in Lemma 10 together with the standard Bernstein inequality (13) imply a version of the empirical Bernstein inequality in [23, Theorem 4] with slightly worse factors. However, the latter inequality cannot be used to derive our main result — we do really require our new inequality to show Theorem 1, since we need to

‘chain’ it to work with samples of length n rather than 1 in a different way. In the next proposition, we will use the following grid \mathcal{G} and distribution π ,

$$\mathcal{G} := \left\{ \frac{1}{\rho}, \dots, \frac{1}{\rho^K} : K := \left\lceil \log_{\rho} \left(\sqrt{\frac{n}{2 \ln \frac{2}{\delta}}} \right) \right\rceil \right\}, \text{ and } \pi = \text{uniform distribution over } \mathcal{G}. \quad (55)$$

for $\rho > 0$. To simplify the presentation, we will use $\rho = 2$ in the next proposition, albeit this may not be the optimal choice.

Proposition 14. *Let \mathcal{G} be as in (55) with $\rho = 2$, and Z, Z_1, \dots, Z_n be i.i.d random variables taking values in $[0, 1]$. Then, for all $\delta \in]0, 1[$, with probability at least $1 - \delta$,*

$$\mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^n Z_i \leq \left(3 \sqrt{\frac{\mathbb{V}_n \cdot \ln \frac{2|\mathcal{G}|}{\delta}}{2n} + \frac{11 \ln \frac{2|\mathcal{G}|}{\delta}}{10n}} \right) \vee \frac{11 \ln \frac{2|\mathcal{G}|}{\delta}}{4n} + \frac{c_{1/2} \cdot \ln \frac{2}{\delta}}{2n},$$

where $\mathbb{V}_n := \frac{1}{n} \sum_{i=1}^n (Z_i - \frac{1}{n} \sum_{j=1}^n Z_j)^2$ is the empirical variance.

Proof. Let $\delta \in]0, 1[$. Applying Lemma 10 to $X_i = Z_i - \mathbb{E}[Z]$, for $i \in [n]$, we get, for all $0 < \eta < 1/2$,

$$\mathbb{E}[Z] - Z_i \triangleleft_{\eta} c_{\eta} \cdot (Z_i - \mathbb{E}[Z])^2, \quad \text{where } c_{\eta} := \eta \cdot \vartheta(\eta). \quad (56)$$

Applying Proposition 7-(b) to chain (56) for $i = 1, \dots, n$, then dividing by n yields

$$\mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^n Z_i \triangleleft_{n\eta} \frac{c_{\eta}}{n} \sum_{i=1}^n (Z_i - \mathbb{E}[Z])^2, \quad (57)$$

$$= c_{\eta} \cdot \mathbb{V}_n + c_{\eta} \cdot \left(\mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^n Z_i \right)^2, \quad (58)$$

where the equality follows from the standard bias-variance decomposition. Let \mathcal{G} and π be as in (55), and let $\hat{\eta} = \hat{\eta}(Z_{\leq n})$ be any random estimator with support on \mathcal{G} . By Proposition 9, a version of (58) with η is replaced by $\hat{\eta}$ and $\ln(|\mathcal{G}|)/(n\hat{\eta})$ added to its RHS also holds. By applying Proposition 6 to this new inequality, we get, with probability at least $1 - \delta$,

$$\mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^n Z_i \leq c_{\hat{\eta}} \cdot \mathbb{V}_n + \frac{\ln \frac{|\mathcal{G}|}{\delta}}{n \cdot \hat{\eta}} + c_{\hat{\eta}} \cdot \left(\mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^n Z_i \right)^2. \quad (59)$$

Now using Hoeffding’s inequality [23, Theorem 3], we also have

$$\left(\mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^n Z_i \right)^2 \leq \frac{\ln \frac{1}{\delta}}{2n}, \quad (60)$$

with probability at least $1 - \delta$. Thus, by combining (59) and (60) via the union bound, we get that, with probability at least $1 - \delta$,

$$\mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^n Z_i \leq \left(c_{\hat{\eta}} \cdot \mathbb{V}_n + \frac{\ln \frac{2|\mathcal{G}|}{\delta}}{n \cdot \hat{\eta}} \right) + \frac{c_{\hat{\eta}} \cdot \ln \frac{2}{\delta}}{2n}. \quad (61)$$

We now use the fact that for all $\eta \in]0, 1/2[$,

$$c_{\eta} = \eta \cdot \vartheta(\eta) \leq \frac{\eta}{2} + \frac{11\eta^2}{20}. \quad (62)$$

Let $\hat{\eta}_* \in [0, +\infty]$ be the un-constrained estimator defined by

$$\hat{\eta}_* := \sqrt{\frac{2 \ln \frac{2|\mathcal{G}|}{\delta}}{\mathbb{V}_n \cdot n}}.$$

Note that by our choice of \mathcal{G} in (55), we always have $\hat{\eta}_* \geq \min \mathcal{G}$. Let $\hat{\eta} \in ([\hat{\eta}_*/2, \hat{\eta}_*] \cap \mathcal{G}) \neq \emptyset$, if $\hat{\eta}_* \leq 1$, and $\hat{\eta} = 1/2$, otherwise. In the first case (i.e. when $\hat{\eta}_* \leq 1$), substituting η for $\hat{\eta} \in$

$([\hat{\eta}_*/2, \hat{\eta}_*] \cap \mathcal{G})$ in the expression between brackets in (61), and using the fact that $\hat{\eta}_*/2 \leq \hat{\eta} \leq \hat{\eta}_*$ and (62), gives

$$c_{\hat{\eta}} \cdot \mathbb{V}_n + \frac{\ln \frac{2|\mathcal{G}|}{\delta}}{\hat{\eta} \cdot n} \leq (1+2) \sqrt{\frac{\mathbb{V}_n \cdot \ln \frac{2|\mathcal{G}|}{\delta}}{2n} + \frac{11 \cdot \ln \frac{2|\mathcal{G}|}{\delta}}{10n}}. \quad (63)$$

Now for the case where $\hat{\eta}_* \geq 1$, we substitute η for $\hat{\eta} = 1/2$ in the expression between brackets in (61), and use (62) and the fact that $1 \leq \hat{\eta}_* = \sqrt{2 \ln(2|\mathcal{G}|/\delta)/(\mathbb{V}_n \cdot n)}$, we get:

$$\begin{aligned} c_{\hat{\eta}} \cdot \mathbb{V}_n + \frac{\ln \frac{2|\mathcal{G}|}{\delta}}{\hat{\eta} \cdot n} &\leq \left(\frac{\hat{\eta}}{2} + \frac{11\hat{\eta}^2}{20} \right) \cdot \mathbb{V}_n + \frac{2 \cdot \ln \frac{2|\mathcal{G}|}{\delta}}{n}, \\ &\leq \left(\frac{\hat{\eta}}{2} + \frac{11\hat{\eta}^2}{20} \right) \cdot \frac{2 \ln \frac{2|\mathcal{G}|}{\delta}}{n} + \frac{2 \cdot \ln \frac{2|\mathcal{G}|}{\delta}}{n}, \quad (\text{due to } \hat{\eta}_* \geq 1) \\ &= \frac{11 \ln \frac{2|\mathcal{G}|}{\delta}}{4n}, \quad (\hat{\eta} = 1/2) \end{aligned} \quad (64)$$

Combining (61), with (63) and (64) yields the desired results. \square