The Annals of Statistics
2019, Vol. 47, No. 5, 2822–2854
https://doi.org/10.1214/18-AOS1766

EXACT LOWER BOUNDS FOR THE AGNOSTIC PROBABLY-APPROXIMATELY-CORRECT (PAC) MACHINE LEARNING MODEL

BY ARYEH KONTOROVICH AND IOSIF PINELIS

Ben-Gurion University and Michigan Technological University

We provide an exact nonasymptotic lower bound on the minimax expected excess risk (EER) in the agnostic probably-approximately-correct (PAC) machine learning classification model and identify minimax learning algorithms as certain maximally symmetric and minimally randomized "voting" procedures. Based on this result, an exact asymptotic lower bound on the minimax EER is provided. This bound is of the simple form $c_\infty/\sqrt{\nu}$ as $\nu \to \infty$, where $c_\infty = 0.16997\dots$ is a universal constant, $\nu = m/d$, m is the size of the training sample and d is the Vapnik–Chervonenkis dimension of the hypothesis class. It is shown that the differences between these asymptotic and nonasymptotic bounds, as well as the differences between these two bounds and the maximum EER of any learning algorithms that minimize the empirical risk, are asymptotically negligible, and all these differences are due to ties in the mentioned "voting" procedures. A few easy to compute nonasymptotic lower bounds on the minimax EER are also obtained, which are shown to be close to the exact asymptotic lower bound $c_\infty/\sqrt{\nu}$ even for rather small values of the ratio $\nu = m/d$. As an application of these results, we substantially improve existing lower bounds on the tail probability of the excess risk. Among the tools used are Bayes estimation and apparently new identities and inequalities for binomial distributions.

1. Introduction. The Probably Approximately Correct (PAC) model aims at providing a clean, plausible and minimalistic abstraction of the supervised learning process [24, 25]. In this paper, we are concerned with the version of this model most commonly appearing in modern literature, *agnostic PAC* [9, 11, 12].

Let $\mathcal{X}$ be an arbitrary nonempty set. The objective is to classify the elements of the set $\mathcal{X}$ into two classes, by attaching a label 1 or -1 to each $x \in \mathcal{X}$. Let $\mathcal{Y} := \{-1, 1\}$, the set of labels. Then a possible classification rule may be identified with a map $h\colon \mathcal{X} \to \mathcal{Y}$, called a *hypothesis*. Usually, hypotheses are restricted to be elements of a specified subset $\mathcal{H}$ of the set $\mathcal{Y}^{\mathcal{X}}$ of all maps of $\mathcal{X}$ to $\mathcal{Y}$; this subset $\mathcal{H}$ is called the *hypothesis class*.

Received June 2016; revised December 2017.

MSC2010 subject classifications. Primary 68T05, 62C20, 62C10, 62C12, 62G20, 62H30; secondary 62G10, 62C20, 91A35, 60C05.

Key words and phrases. PAC learning theory, classification, generalization error, minimax decision rules, Bayes decision rules, empirical estimators, binomial distribution.

It is assumed that there exists a true (but unknown to us) probability distribution, say D, on the set $\mathcal{X} \times \mathcal{Y}$ of all pairs (x, y) with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. To avoid tedious matters of measurability, let us just assume that the set $\mathcal{X}$ is finite.

In the agnostic PAC model, considered in this paper, it is assumed that the distribution D may be of completely arbitrary form, and the only information about it is provided to us by the "sample" values of a *labeled sample* $(X_1^D, Y_1^D), \dots, (X_m^D, Y_m^D)$ of m independent copies of a random pair (X^D, Y^D); here and in what follows, the superscript indicates the distribution of the random pair.

The classification error probability for a hypothesis $h \in \mathcal{H}$ is

$$\text{err}(h, D) := \mathsf{P}\big(h(X^D) \neq Y^D\big). \tag{1.1}$$

It should be clear that the least possible error probability

$$\underset{\min}{\text{err}}(D) := \underset{\min,\mathcal{H}}{\text{err}}(D) := \min_{h \in \mathcal{H}} \text{err}(h, D) \tag{1.2}$$

will usually be strictly greater than 0, even when the true distribution D is known.

In the agnostic PAC model, considered here, the only information about the unknown distribution D is provided by the values of the sequence of m independent random pairs

$$Z_m^D := \big((X_1^D, Y_1^D), \dots, (X_m^D, Y_m^D)\big). \tag{1.3}$$

Therefore, the available "learning" strategies are the mappings

$$L\colon (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H},$$

called *learning algorithms*.

Let h_D denote any minimizer of $\text{err}(h, D)$ over $h \in \mathcal{H}$. Of course, h_D is unknown, since the distribution D is unknown. However, it may be reasonable to use the plug-in estimator $h_{\hat{D}_m}$ of h_D, obtained by substituting for D the empirical distribution $\hat{D}_m = \hat{D}_m(z_m)$ based on a "realization"

$$z_m := \big((x_1, y_1), \dots, (x_m, y_m)\big) \in (\mathcal{X} \times \mathcal{Y})^m \tag{1.4}$$

of the "random sample" Z_m from the distribution D. That is,

$$h_{\hat{D}_m} = L_{\mathsf{ERM}}(Z_m^D),$$

where L_{ERM} is an empirical risk minimizer, that is, any learning algorithm such that for each given sequence $z_m \in (\mathcal{X} \times \mathcal{Y})^m$, the corresponding value $L_{\mathsf{ERM}}(z_m)$ of L_{ERM} is a minimizer in $h \in \mathcal{H}$ of the "empirical risk"

$$\text{err}(h, \hat{D}_m) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{I}\big\{h(x_i) \neq y_i\big\}.$$

Such a minimizer need not be unique, and so, the "empirical minimization" learning algorithm L_{ERM} does not have to be unique.

A nontrivial question to ask here is how well the empirical risk minimizer $h_{\hat{D}_m}$ performs compared to the best possible hypothesis, h_D, that is, how large the *excess risk* $\Delta(h_{\hat{D}_m}, D)$ is, where

$$\Delta(h, D) := \operatorname{err}(h, D) - \operatorname*{err}_{\min}(D) = \operatorname{err}(h, D) - \operatorname{err}(h_D, D). \tag{1.5}$$

This question has been to a large extent resolved. In particular, Theorem 4.9 from [1] (slightly restated here) provides the following upper bound on the tail probabilities for the excess risk.

THEOREM A. *There is a universal real constant $c > 0$ such that for all finite sets $\mathcal{X}$, all distributions D on $\mathcal{X} \times \mathcal{Y}$, all sample sizes m, and all hypothesis classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of VC dimension d, we have*

$$\mathsf{P}\big(\Delta\big(L_{\mathsf{ERM}}(Z_m^D), D\big) \geq cu\big) \leq \exp\big\{-\big(mu^2 - d\big)_+\big\} \tag{1.6}$$

for all real $u \geq 0$, where $z_+ := 0 \vee z$ for real z.

See [1] for an account of the intermediate steps leading up to the highly nontrivial result presented in Theorem A; milestones here include the seminal paper [25] by Vapnik and Chervonenkis, followed, notably, by work of Talagrand [22], Haussler [10] and Long [15].

Recall that the VC dimension (i.e., the Vapnik–Chervonenkis dimension) of a set $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is the largest nonnegative integer k such that there is a subset of $\mathcal{X}$ of cardinality k that is shattered by $\mathcal{H}$; and a subset $\mathcal{X}_0$ of $\mathcal{X}$ is said to be shattered by $\mathcal{H}$ if the set of the restrictions to $\mathcal{X}_0$ of all the functions $h \in \mathcal{H}$ coincides with the entire set $\mathcal{Y}^{\mathcal{X}_0}$ of all functions from $\mathcal{X}_0$ to $\mathcal{Y}$.

In what follows, d will always denote $\operatorname{VC}(\mathcal{H})$, the VC dimension of $\mathcal{H}$. The case $d = 0$ may occur only if the cardinality of $\mathcal{H}$ is at most 1, so that there is at most one hypothesis to choose. This trivial case will be excluded in the sequel; that is, we shall assume that $d = 1$. Then, in particular, one can introduce the fundamental ratio

$$\nu := m/d \tag{1.7}$$

of the sample size m to the VC dimension d.

Lower bounds matching, up to constant factors, the upper bound given in Theorem A are also known. The one with the apparently best currently known numerical constants was given in [1], Theorem 5.2, which can be restated as follows.

THEOREM B. *If $\nu = m/d \geq 64^2/320 = 12.8$, then for any finite set $\mathcal{X}$, any hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of VC dimension d, and any learning algorithm $L \colon (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$, there is a distribution D on $\mathcal{X} \times \mathcal{Y}$ such that*

$$\mathsf{P}\left(\Delta\big(L(Z_m^D), D\big) > \frac{1}{\sqrt{320\nu}}\right) \geq \frac{1}{64}. \tag{1.8}$$

This lower bound is also the culmination of a notable historical development [6, 20, 25], detailed in [1].

REMARK 1.1. In [19], Section 28.2.2, a much better constant factor, $1/8$, was claimed in place of 320 in (1.8). However, there is a mistake in the calculation; the actual value of the constant furnished by the proof is 512.

Introduce the *expected excess risk* (EER)

$$\mathfrak{R}(L, D) := \mathfrak{R}_m(L, D) := \mathsf{E}\,\Delta\big(L\big(Z_m^D\big), D\big). \tag{1.9}$$

Let $\mathcal{D} := \mathcal{D}_{\mathcal{X}}$ and $\mathcal{L} := \mathcal{L}_{\mathcal{X};m,\mathcal{H}}$ denote, respectively, the set of all distributions on $\mathcal{X} \times \mathcal{Y}$ and the set of all learning algorithms $L\colon (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$; recall here that $\mathcal{Y} = \{-1, 1\}$. Let then

$$\begin{aligned} c_{m,d}^{\mathsf{UB}} &:= \sqrt{m/d} \sup_{\mathcal{X}} \sup_{\mathrm{VC}(\mathcal{H})=d} \inf_{L \in \mathcal{L}_{\mathcal{X};m,\mathcal{H}}} \sup_{D \in \mathcal{D}_{\mathcal{X}}} \mathfrak{R}_m(L, D), \\ c_{m,d}^{\mathsf{LB}} &:= \sqrt{m/d} \inf_{\mathcal{X}} \inf_{\mathrm{VC}(\mathcal{H})=d} \inf_{L \in \mathcal{L}_{\mathcal{X};m,\mathcal{H}}} \sup_{D \in \mathcal{D}_{\mathcal{X}}} \mathfrak{R}_m(L, D), \end{aligned} \tag{1.10}$$

where $\sup_{\mathcal{X}}$ and $\inf_{\mathcal{X}}$ are taken over all finite sets $\mathcal{X}$, and $\sup_{\mathrm{VC}(\mathcal{H})=d}$ and $\inf_{\mathrm{VC}(\mathcal{H})=d}$ are taken over all hypothesis classes $\mathcal{H}$ of VC dimension d. The quantity $\inf_{L\in\mathcal{L}} \sup_{D\in\mathcal{D}} \mathfrak{R}_m(L, D)$ may be referred to as the *minimax* EER.

Integrating both sides of inequality (1.6) in $u \geq 0$, one sees that

$$c^{\mathsf{UB}} := \sup_{m,d} c_{m,d}^{\mathsf{UB}} \leq \sup_{m,d} \sqrt{m/d} \sup_{\mathcal{X}} \sup_{\mathrm{VC}(\mathcal{H})=d} \sup_{D \in \mathcal{D}_{\mathcal{X}}} \mathfrak{R}_m(L_{\mathsf{ERM}}, D) < \infty, \tag{1.11}$$

where $\sup_{m,d}$ is taken over all natural m and d; an exact calculation of c^{UB} seems to be beyond the reach of current methods.

It is also clear that inequality (1.8) implies

$$c_{\nu \geq 12.8}^{\mathsf{LB}} > \frac{1}{64\sqrt{320}} = 0.000873\ldots > 0, \tag{1.12}$$

where $c_{\nu\geq\nu_*}^{\mathsf{LB}} := \inf\{c_{m,d}^{\mathsf{LB}} \colon m/d \geq \nu_*\}$. for any real $\nu_* > 0$. A remarkable fact that follows from (1.11) and (1.12) is that

$$0 < \liminf_{m/d\to\infty} c_{m,d}^{\mathsf{LB}} \leq \limsup_{m/d\to\infty} c_{m,d}^{\mathsf{UB}} < \infty;$$

that is, the upper and lower bounds on the minimax EER are of the same order of magnitude. Establishing an appropriate lower bound on the EER, $\mathfrak{R}(L, D)$, was the crucial part of the proof of Theorem B.

A few words on the organization of the rest of this paper: The main results are stated and discussed in Section 2. All necessary proofs are given in Section 3, with more technical parts deferred further, to Appendices A–B.

An index of symbols used in this paper nonlocally is given in Table 1, which lists the places where the selected symbols are first introduced and, for a few of the symbols, the places where those symbols are generalized, specialized or otherwise modified.

TABLE 1
Notation. The places where the symbols are first introduced are underlined

Symbol	**Brief description**	**Appears in/on**
$B(m,d)$	expression for $\inf_{L\in\mathcal{L}_{\mathsf{rand}}}\sup_D \mathfrak{R}_m(L,D)$	(2.13)
$B_0(m,d)$	lower bound on $B(m,d)$	(2.21)
$B_1(\nu)$	lower bound on $B_0(m,d)$	(2.26)
$B_2(\nu)$	lower bound on $B_1(\nu)$	(2.28)
$\tilde{B}_2(\nu)$, $\hat{B}_2(\nu)$	lower bounds on $B_2(\nu)$	(2.30), (2.32)
$\mathrm{bayes}(k,b)$	Bayes risk for $d=1$	(2.14); (3.2)
$\underset{\smile}{\mathrm{bayes}}(\kappa,b)$	convex minorant of $\mathrm{bayes}(k,b)$	Proposition 2.8
$b\in[-1,1]$	$\mathcal{Y}$-bias for $\mathcal{X}=\{1\}$	symbols $s_k(b),\ldots$
$\beta(x)$	conditional $\mathcal{Y}$-bias at $x\in\mathcal{X}$	(2.3)
$c^{\mathsf{UB}}_{m,d}$ $[c^{\mathsf{LB}}_{m,d}]$	$\nu\times$ (exact upper [lower] bound on the minimax EER)	(1.10)
$c_\infty=0.16997\ldots$	limit value of $c^{\mathsf{LB}}_{m,d}$	(2.1)
c_ν, $\tilde{c}_\nu$	close lower bounds on $c^{\mathsf{LB}}_{m,d}$	(2.29), (2.30)
C_i		(2.29), (A.12)
D, $D_{p,\beta}$	distribution on $\mathcal{X}\times\mathcal{Y}$	page 2822, page 2829
$d:=\mathrm{VC}(\mathcal{H})$	VC dimension of $\mathcal{H}$	page 2824
$\Delta(h,D)$	excess risk	(1.5)
$\mathrm{err}(h,D)$	error probability	(1.1)
$\mathrm{err}_{\mathrm{min}}(h,D)$	minimum error probability	(1.2)
$h\colon\mathcal{X}\to\mathcal{Y}$	hypothesis	page 2822
$\mathcal{H}$	hypothesis class	page 2822, page 2828
$\mathrm{I}\{\cdot\}$	indicator function	below (2.6)
L	learning algorithm (l.a.)	page 2823
L_{ERM}, L^*_{ERM}	empirical risk minimizer	page 2823, (2.7), (2.11)
$\mathcal{L}$	set of all nonrandomized l.a.'s	page 2825, page 2830
$\mathcal{L}_{\mathsf{rand}}$	set of all randomized l.a.'s	page 2830
m	labeled sample size	page 2822
N	binomial r.v. w/ parameters m, $1/d$	Theorem 2.5
N^p_x	cardinality of the set $\{i\colon X_i=x\}$	Theorem 2.2; (3.5)
$\nu:=m/d$	fundamental ratio	(1.7)
p	$\mathcal{X}$-marginal of D	(2.3)
$\mathfrak{R}(L,D)$, $\mathfrak{R}(L;p,\beta)$	expected excess risk (EER)	(1.9), (2.10)
sgn	modified sign function	below (2.6)
$s_k(b)$		(2.15)
V^b_k, Y^b_i		(2.16)
$V^{p,\beta}_x$	"vote balance" at $x\in\mathcal{X}$	(2.20)
(X^D_i,Y^D_i), $(X^p_i,Y^{p,\beta}_i)$	labeled sample items	page 2822, page 2829
$\mathcal{X}$	set of objects to classify	page 2822, page 2828
$\mathcal{Y}=\{-1,1\}$	set of classification labels	page 2822
Z^D_m, $Z^{p,\beta}_m$	labeled sample	(1.3), (2.4)
$z_*=0.75179\ldots$	maximizer of $\frac{z}{2}(1-\mathrm{erf}(z/\sqrt{2}))$	(2.27)

2. Results: Statements and discussion. In this paper, we present optimal lower bounds on the minimax EER, which cannot be further improved. Our main result is Theorem 2.2, which provides an expression of the exact, nonasymptotic lower bound on the minimax EER. This expression is in terms of a certain function $\mathrm{bayes}(k, b)$, which is the Bayes risk for $d = 1$. We show (in Proposition 2.8) that $\mathrm{bayes}(k, b)$ has a certain convexity property with respect to k. Further important properties of the function bayes, based on certain apparently novel identities and inequalities for binomial distributions, are presented in Appendix A. Thus, the expression of the non-asymptotic lower bound on the minimax EER given in Theorem 2.2 becomes amenable to high-precision analysis. (Implicitly, the function bayes is present in [3], but there it was bounded via Pinsker's inequality, which yields suboptimal results.)

In particular, based on Theorem 2.2 and the mentioned analysis of the function bayes, we determine (in Theorem 2.4) the asymptotics of the just mentioned exact lower bound:

$$c^{\mathsf{LB}}_{m,d} \to c_\infty := \max_{z>0} \frac{z}{2}\bigl(1 - \mathrm{erf}(z/\sqrt{2})\bigr) = 0.16997\ldots \tag{2.1}$$

whenever m and d vary in such a way that $\nu = m/d \to \infty$; here, as usual, erf denotes the Gauss error function, given by the formula $\mathrm{erf}(u) := \frac{2}{\sqrt{\pi}} \int_0^u e^{-t^2}\,dt$ for real $u \geq 0$.

It should be noted that in Theorem 2.2 randomization of learning algorithms is allowed; however, it will also be shown (in Theorem 2.4) that the effect of this randomization is asymptotically negligible and is entirely explained by ties in a certain "voting" procedure.

Theorems 2.5, 2.9, 2.11 and Proposition 2.13 present, for finite m and d, tractable lower bounds on $c^{\mathsf{LB}}_{m,d}$; Theorem 2.16 then shows that all these lower bounds on $c^{\mathsf{LB}}_{m,d}$, as well as $c^{\mathsf{LB}}_{m,d}$ itself, converge to the limit constant c_∞ in (2.1). Moreover, it is shown (see Remarks 2.7 and 2.14, and Figure 2) that these lower bounds on $c^{\mathsf{LB}}_{m,d}$, as well as $c^{\mathsf{LB}}_{m,d}$ itself, are actually close to the limit value c_∞ even for rather small values of $\nu = m/d$.

The above discussion suggests a sense of completion in the area of lower bounds for the PAC model. However, results and techniques presented here may be used elsewhere. In fact, they already found an application in [14], Theorem 7.1, where existing lower bounds were not sufficiently delicate for the desired parameter regime.

In this paper, we apply our lower bounds on the EER to obtain substantial improvements of the existing lower bounds on the tail probability of the excess risk, as follows:

THEOREM 2.1. (i) *Keeping the constants* 12.8 *and* 320 *in Theorem* B *in place, one can improve the lower bound* $\frac{1}{64} \approx 0.0156$ *on the tail probability in* (1.8) *to* 0.238.

(ii) *Keeping the constants* 12.8 *and* $\frac{1}{64}$ *in Theorem* B *in place*, *one can improve the constant* 320 *in* (1.8) *to* 41.3.

(iii) *If the restriction* $\nu \geq 12.8$ *in Theorem* B *is relaxed to* $\nu \geq 3$, *then the improved values* 0.238 *and* 41.3 *of the constants get only slightly worse*: 0.227 *and* 49.6, *respectively.*

To state our results, let us introduce some additional notation and conventions to be used in the sequel.

Let $0^0 := 1$.

For any α and ω in $\mathbb{Z} \cup \{\infty\}$, let $\overline{\alpha, \omega} := \{i \in \mathbb{Z} \colon \alpha \leq i \leq \omega\}$. For any $m \in \overline{0, \infty}$, let $[m] := \overline{1, m}$. In particular, $[0] = \varnothing$.

As usual, for any two sets S and T, let S^T denote the set of all maps from T to S.

For any set A and any $k \in \overline{0, \infty}$ we identify the k-tuples $v = (v_1, \dots, v_k) \in A^k$ with functions $v \colon [k] \to A$, by the formula $v(x) := v_x$ for all $x \in [k]$; thus, we identify the set A^k of k-tuples with the set $A^{[k]}$ of functions. So, we use the notation $v(x)$ and v_x interchangeably. We shall also identify a function with its graph.

As usual, the sum of the empty family of elements of a linear space is defined as the zero element of that space.

The new results obtained in this paper all concern the lower bound $c^{\mathsf{LB}}_{m,d}$, defined in (1.10), on the minimax EER times the factor $\sqrt{m/d}$, including the limit behavior of $c^{\mathsf{LB}}_{m,d}$ as $m/d \to \infty$.

It is not hard to show (see Appendix B for details) that the defining expression for $c^{\mathsf{LB}}_{m,d}$ in (1.10) can be simplified as follows:

$$c^{\mathsf{LB}}_{m,d} = \sqrt{m/d} \inf_L \sup_D \mathfrak{R}_m(L, D), \tag{2.2}$$

where from now on it will be assumed (unless otherwise specified) that

$$\boxed{\mathcal{X} = [d], \qquad \mathcal{H} = \mathcal{Y}^{\mathcal{X}} = \{-1, 1\}^{[d]}, \qquad \inf_L := \inf_{L \in \mathcal{L}_{[d];m,\mathcal{Y}^{[d]}}}, \qquad \sup_D := \sup_{D \in \mathcal{D}_{[d]}},}$$

so that $\inf_L$ is taken over all learning algorithms $L \colon ([d] \times \mathcal{Y})^m \to \mathcal{Y}^{[d]}$ and $\sup_D$ is taken over all distributions D on $[d] \times \mathcal{Y}$. Accordingly, from now on we shall use $\mathcal{X}$ and $\mathcal{H}$ interchangeably with $[d]$ and $\mathcal{Y}^{\mathcal{X}} = \{-1, 1\}^{[d]}$, respectively.

Note next that any distribution D on $\mathcal{X} \times \mathcal{Y}$ is completely characterized by the two maps, say $p = p_D \colon \mathcal{X} \to [0, 1]$ and $\beta = \beta_D \colon \mathcal{X} \to [-1, 1]$, such that

$$\mathsf{P}\big(X^D = x, Y^D = y\big) = D\big(\{(x, y)\}\big) = p(x)\left(\frac{1}{2} + \frac{y\beta(x)}{2}\right) \tag{2.3}$$

for $x \in \mathcal{X} = [d]$ and $y \in \mathcal{Y} = \{-1, 1\}$. Clearly then, one must have $p_D(x) = \mathsf{P}(X^D = x)$ for all $x \in \mathcal{X}$ and $\beta_D(x) = 2\,\mathsf{P}(Y^D = 1 | X^D = x) - 1$ for all $x \in \mathcal{X}$ with $p_D(x) \neq 0$; if $p_D(x) = 0$ for some $x \in \mathcal{X}$, then the value of $\beta_D(x)$ can be chosen arbitrarily in $[-1, 1]$, So, the distribution of the random variable (r.v.) X^D

is completely characterized by the map $p = p_D$. Therefore, in what follows let us write $D = D_{p,\beta}$ if $p_D = p$ and $\beta_D = \beta$, and, in the case when $D = D_{p,\beta}$, let us simply write X^p, $Y^{p,\beta}$, X_i^p, $Y_i^{p,\beta}$ instead of X^D, Y^D, X_i^D, Y_i^D (respectively), assuming that the random pairs $(X_1^p, Y_1^{p,\beta}), \dots, (X_m^p, Y_m^{p,\beta})$ are independent copies of the random pair $(X^p, Y^{p,\beta}) = (X^D, Y^D)$; let us then also write

$$Z_m^{p,\beta} := Z_m^D := \big((X_1^p, Y_1^{p,\beta}), \dots, (X_m^p, Y_m^{p,\beta})\big) \tag{2.4}$$

(cf. (1.3)).

Take next any $h \in \mathcal{H} = \{-1, 1\}^{[d]}$. It is well known (see, e.g., [5], page 10) that the function $h_\beta \in \{-1, 1\}^{[d]}$ given by the formula

$$h_\beta(x) := \operatorname{sgn} \beta_x \tag{2.5}$$

for $x \in [d]$ is a minimizer of $\operatorname{err}(h, D_{p,\beta})$ over all $h \in \{-1, 1\}^{[d]}$, and the excess risk (relative to $D_{p,\beta}$) of h over h_β is

$$\begin{aligned}\Delta(h, D_{p,\beta}) &= \operatorname{err}(h, D_{p,\beta}) - \operatorname{err}(h_\beta, D_{p,\beta}) \\ &= \sum_{x=1}^d p_x |\beta_x| \,\mathrm{I}\big\{h(x) \neq h_\beta(x)\big\},\end{aligned} \tag{2.6}$$

where

$$\operatorname{sgn} u := 2\,\mathrm{I}\{u \geq 0\} - 1$$

for real u and $\mathrm{I}\{\cdot\}$ is the indicator function.

Replacing now the unknown true distribution $D = D_{p,\beta}$ by the empirical distribution $\hat{D}_m = \hat{D}_m((x_1, y_1), \dots, (x_m, y_m))$ for $((x_1, y_1), \dots, (x_m, y_m)) =: z \in (\mathcal{X} \times \mathcal{Y})^m$, one sees that a function $h \in \{-1, 1\}^{[d]}$ is a minimizer of $\operatorname{err}(h, \hat{D}_m)$ for the given "sample" z if and only if $h(x) = \operatorname{sgn} \widehat{p\beta}_x$ for all $x \in \mathcal{X}$ such that $\widehat{p\beta}_x \neq 0$, where $\widehat{p\beta}_x := \frac{1}{m} \sum_{i=1}^m y_i \,\mathrm{I}\{x_i = x\}$; if $\widehat{p\beta}_x = 0$ for some $x \in \mathcal{X}$, then the value $h(x)$ of a minimizer h (of $\operatorname{err}(h, \hat{D}_m)$) at this point x can be chosen arbitrarily in the set $\{-1, 1\}$. Thus, all the learning algorithms L_{ERM} that are minimizers of the empirical risk are given by the formula

$$L_{\mathsf{ERM}}(z_m)(x) := L_{m,d;\mathsf{ERM}}(z_m)(x) \begin{cases} := \operatorname{sgn} v_x & \text{if } v_x \neq 0, \\ \in \{-1, 1\} & \text{if } v_x = 0 \end{cases} \tag{2.7}$$

for all $z_m \in (\mathcal{X} \times \mathcal{Y})^m$ and $x \in \mathcal{X} = [d]$, where

$$v_x := v_x(z_m) := \sum_{i=1}^m y_i \,\mathrm{I}\{x_i = x\} = m\, \widehat{p\beta}_x. \tag{2.8}$$

Formula (2.7) states that the empirical risk is minimized when the value $y \in \{-1, 1\}$ assigned by the learning algorithm at point x based on the "sample" z_m is decided by the majority vote $v_x = v_x(z_m)$ "at x," with the "voting" restricted to the pairs (x_i, y_i) with $x_i = x$; if there is a tie (no majority) at x, then a value $y \in \{-1, 1\}$ at x is chosen arbitrarily.

To decrease the risk and also be able to fully use the power of decision theory, one may randomize learning algorithms. A convenient way to define such an algorithm L is to allow its value (which is a function in $\mathcal{H}$) to depend, not only on the nonrandom "sample" $z_m = ((x_1, y_1), \dots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ as in (1.4), but also on the value u of another r.v., say U, which is (say) uniformly distributed on the interval $[-1, 1]$ and independent of the random "sample" $Z_m^D = ((X_1^D, Y_1^D), \dots, (X_m^D, Y_m^D))$ as in (1.3). Thus, a randomized learning algorithm L will be understood as a map from $(\mathcal{X} \times \mathcal{Y})^m \times [-1, 1]$ to $\mathcal{H}$.

Let $\mathcal{L}_{\mathsf{rand}} = \mathcal{L}_{\mathsf{rand},d}$ and $\mathcal{L} = \mathcal{L}_d$ denote, respectively, the set of all randomized learning algorithms and the set of all nonrandomized ones. The definition (1.9) of the EER (for $L \in \mathcal{L}$) is naturally extended as follows:

$$\mathfrak{R}(L, D) := \mathfrak{R}_m(L, D) := \mathsf{E}\,\Delta\big(L\big(Z_m^D, U\big), D\big) \tag{2.9}$$

for $L \in \mathcal{L}_{\mathsf{rand}}$; it then follows by (2.6) that

$$\begin{aligned} \mathfrak{R}_m(L; p, \beta) &:= \mathfrak{R}_m(L, D_{p,\beta}) \\ &= \sum_{x=1}^{d} p_x |\beta_x| \, \mathsf{P}\big(L\big(Z_m^{p,\beta}, U\big)(x) \neq h_\beta(x)\big). \end{aligned} \tag{2.10}$$

Of particular importance will be the following "maximally symmetric" and "minimally randomized" version of the learning algorithms L_{ERM} that are minimizers of the empirical risk (cf. (2.7)):

$$\begin{aligned} L^*_{\mathsf{ERM}}(z_m, u)(x) &:= L^*_{d,\mathsf{ERM}}(z_m, u)(x) \\ &:= \begin{cases} \operatorname{sgn} v_x & \text{if } v_x \neq 0, \\ y_{i_x} & \text{if } v_x = 0 \text{ but } n_x \neq 0, \\ \operatorname{sgn} u & \text{if } n_x = 0 \end{cases} \end{aligned} \tag{2.11}$$

for $(z_m, u) \in (\mathcal{X} \times \mathcal{Y})^m \times [-1, 1] = ([d] \times \{-1, 1\})^m \times [-1, 1]$, where

$$\begin{aligned} n_x &:= n_x(z_m) := \sum_{i=1}^{m} \mathrm{I}\{x_i = x\} \quad \text{and} \\ i_x &:= i_x(z_m) := \min\big\{i \in [m] \colon x_i = x\big\}. \end{aligned} \tag{2.12}$$

That is, the choice of the value of $L^*_{\mathsf{ERM}}(z_m, u)(x)$ in $\mathcal{Y} = \{-1, 1\}$ is decided by the majority vote "at x" if there is a majority there; otherwise, the value $L^*_{\mathsf{ERM}}(z_m, u)(x)$ is the same as that of the first voter that appeared "at x" if any one did; finally, if no one arrived to vote "at x," then the value is decided by a flip of a fair coin, the flip being independent of any voters. Thus, randomization according to the learning algorithm L^*_{ERM} occurs only if no one shows up for voting at some location $x \in \mathcal{X}$. Yet, this minimal (and, one may argue, quite natural) randomization is enough to make L^*_{ERM} a winner (i.e., a minimax learning algorithm) against all randomized (and nonrandomized) learning algorithms. A precise formulation of this thesis is contained in

THEOREM 2.2. *Take any* $m \in \overline{0,\infty}$. *Then*

$$\inf_{L\in\mathcal{L}_{\mathsf{rand}}} \sup_D \mathfrak{R}_m(L,D) = \sup_D \mathfrak{R}_m\big(L^*_{\mathsf{ERM}}, D\big)$$

$$= B(m,d) := \sup_{p,\beta} \sum_{x=1}^d p_x|\beta_x| \, \mathsf{E}\,\mathrm{bayes}\big(N_x^p, |\beta_x|\big), \tag{2.13}$$

where $\sup_{p,\beta}$ *is taken over all pairs of functions* $p \in [0,1]^{[d]}$ *such that* $\sum_{x=1}^d p_x = 1$ *and* $\beta \in [-1,1]^{[d]}$, N_x^p *is a r.v. with the binomial distribution with parameters* m *and* p_x,

$$\mathrm{bayes}(k,b) := \frac{1}{2}\big(1 - s_k(b)\big), \tag{2.14}$$

$$s_k(b) := \big|\mathsf{P}\big(V_k^b > 0\big) - \mathsf{P}\big(V_k^{-b} > 0\big)\big|, \tag{2.15}$$

$$V_k^b := Y_1^b + \cdots + Y_k^b, \tag{2.16}$$

and the Y_i^b*'s are i.i.d. r.v.'s with* $\mathsf{P}(Y_i^b = 1) = \frac{1+b}{2}$ *and* $\mathsf{P}(Y_i^b = -1) = \frac{1-b}{2}$ $[= 1 - \mathsf{P}(Y_i^b = 1)]$, *for* $k \in \overline{0,\infty}$ *and* $b \in [-1,1]$. *Moreover, for each pair of functions* p *and* β *as described above*,

$$\mathfrak{R}_m\big(L^*_{\mathsf{ERM}}; p,\beta\big) = \mathfrak{R}_m\big(L^*_{\mathsf{ERM}}, D_{p,\beta}\big) = \sum_{x=1}^d p_x|\beta_x| \, \mathsf{E}\,\mathrm{bayes}\big(N_x^p, |\beta_x|\big), \tag{2.17}$$

which does not depend on $\operatorname{sgn}\beta := (\operatorname{sgn}\beta_1, \dots, \operatorname{sgn}\beta_d)$.

The use of the symbol bayes in (2.13) is a reflection of the fact that the minimax learning algorithm L^*_{ERM} is a Bayes one with respect to a certain prior distribution on the set of all distributions D on $\mathcal{X}\times\mathcal{Y}$; see the beginning of the proof of Theorem 2.2 in Section 3 for details on this. Formula (2.17) means that the learning algorithm L^*_{ERM} has an important risk-equalizing property, which actually makes the Bayes decision rule L^*_{ERM} minimax; cf., for example, Theorem 3 and Lemma 1 in [7], Section 2.11.

REMARK 2.3. It is clear from (2.14)–(2.15) that bayes $\le \frac{1}{2}$. Hence, by (2.13), $\inf_{L\in\mathcal{L}_{\mathsf{rand}}} \sup_D \mathfrak{R}_m(L,D) \le \frac{1}{2}$.

It turns out, as may be expected, that the effect of the randomization of learning algorithms is asymptotically negligible whenever $\nu = m/d \to \infty$; that is, the difference $\inf_{L\in\mathcal{L}} \sup_D - \inf_{L\in\mathcal{L}_{\mathsf{rand}}} \sup_D$ is asymptotically negligible compared with the "nonrandomized" minimax EER $\inf_{L\in\mathcal{L}} \sup_D$. Moreover, all the learning algorithms of the form L_{ERM} as in (2.7) that are minimizers of the empirical risk are asymptotically minimax. These facts—along with the asymptotics of the minimax risk—are presented in the following.

THEOREM 2.4. *For each pair* (m, d) *of natural numbers, choose any learning algorithm of the form* $L_{m,d;\mathsf{ERM}}$, *as in* (2.7). *Then*

$$\begin{aligned}\frac{c_\infty}{\sqrt{m/d}} &\sim \inf_{L\in\mathcal{L}_{\mathsf{rand}}} \sup_D \mathfrak{R}_m(L, D) \le \inf_{L\in\mathcal{L}} \sup_D \mathfrak{R}_m(L, D)\\ &\le \sup_D \mathfrak{R}_m(L_{m,d;\mathsf{ERM}}, D) \sim \frac{c_\infty}{\sqrt{m/d}}\end{aligned} \tag{2.18}$$

whenever $m/d \to \infty$, *where* $c_\infty = 0.16997\ldots$ *as in* (2.1). *Moreover,*

$$\begin{aligned}0 &\le \sup_D \mathfrak{R}_m(L_{m,d;\mathsf{ERM}}, D) - \inf_{L\in\mathcal{L}_{\mathsf{rand}}} \sup_D \mathfrak{R}_m(L, D)\\ &\le \frac{1}{2}\sup_{p,\beta} \sum_{x=1}^d p_x|\beta_x|\,\mathsf{P}(V_x^{p,\beta} = 0) = O\left(\frac{1}{m/d}\right) = o\left(\frac{1}{\sqrt{m/d}}\right),\end{aligned} \tag{2.19}$$

again whenever $m/d \to \infty$, *where*

$$V_x^{p,\beta} := \sum_{i=1}^m Y_i^{p,\beta}\, \mathrm{I}\{X_i^p = x\}, \tag{2.20}$$

the vote "balance" at x *based on the random "sample"* $Z_m^{p,\beta}$ *as in* (2.4).

Here, as usual, the asymptotic equivalence $A \sim B$ means $A/B \to 1$.

Display (2.19) shows that the (asymptotically negligible) pairwise differences between (i) the minimax EER $\inf_{L\in\mathcal{L}} \sup_D \mathfrak{R}_m(L, D)$, (ii) its "randomized" version $\inf_{L\in\mathcal{L}_{\mathsf{rand}}} \sup_D \mathfrak{R}_m(L, D)$, and (iii) the maximum risk $\sup_D \mathfrak{R}_m(L_{m,d;\mathsf{ERM}}, D)$ of any empirical-risk-minimizing learning algorithms of the form L_{ERM} are entirely explained by ties in the mentioned "voting," when the "no-majority" event $V_x^{p,\beta} = 0$ occurs for at least one $x \in \mathcal{X} = [d]$.

It is obvious from (2.13) that

$$B(m, d) \ge \sum_{x=1}^d \frac{1}{d} b\,\mathsf{E}\,\mathrm{bayes}(N_x, b) = b\,\mathsf{E}\,\mathrm{bayes}(N_1, b)$$

for any $b \in [0, 1]$, where N_x stands for N_x^p with $p_x = \frac{1}{d}$ for all $x \in [d]$. Thus, in view of (2.18), one immediately obtains the following.

THEOREM 2.5.

$$\inf_{L\in\mathcal{L}}\sup_D \mathfrak{R}_m(L,D) \ge \inf_{L\in\mathcal{L}_{\mathsf{rand}}}\sup_D \mathfrak{R}_m(L,D)$$
$$\ge B_0(m,d) := \sup_{b\in[0,1]} b\,\mathsf{E}\,\mathrm{bayes}(N,b), \tag{2.21}$$

where N is a binomial r.v. with parameters m and $1/d$.

Recall (2.4) and let

$$D_\beta := D_{p,\beta} \quad\text{and}\quad Z_m^\beta := Z_m^{p,\beta} \qquad \text{when } p_x = \frac{1}{d} \text{ for all } x\in\mathcal{X} = [d]. \tag{2.22}$$

THEOREM 2.6. *For any $b\in[0,1]$,*

$$\inf_{L\in\mathcal{L}_{\mathsf{rand}}} \frac{1}{2^d}\sum_{\beta\in\{-b,b\}^{[d]}} \mathfrak{R}(L,D_\beta) = b\,\mathsf{E}\,\mathrm{bayes}(N,b). \tag{2.23}$$

As we shall see, Theorem 2.6 follows immediately from the proof of Theorem 2.2. On the other hand, Theorem 2.6 could be viewed as a refinement of Theorem 2.5, because clearly $\frac{1}{2^d}\sum_{\beta\in\{-b,b\}^{[d]}}\mathfrak{R}(L,D_\beta) \le \sup_D \mathfrak{R}(L,D)$ for any L. Even though the refinement is slight, Theorem 2.6 will be useful, in particular, in the proof of Theorem 2.1.

REMARK 2.7. Note that, by (2.14)–(2.15), $\mathrm{bayes}(k,b)$ is a polynomial in b of degree $\le k$. Hence, $b\,\mathsf{E}\,\mathrm{bayes}(N,b)$ is a polynomial in b of degree $\le m+1$, and so, the lower bound $B_0(m,d)$ in (2.21) is an algebraic number, which is not hard to compute unless m is too large. For instance, for $c_0(m,d) := B_0(m,d)\sqrt{m/d}$ we find

$$c_0(5,2) = 0.16757\ldots,$$
$$c_0(50,20) = 0.17467\ldots \quad\text{and} \tag{2.24}$$
$$c_0(50,2) = 0.16968\ldots$$

(with the execution times in Mathematica about 0.02 sec, 1.4 sec and 1 sec, resp.). One may note that, even for such a rather small value $5/2 = 50/20 = 2.5$ of $\nu = m/d$, the values of $c_0(m,d)$ are close to the limit value $c_\infty = 0.16997\ldots$; cf. (2.1) and (2.18). However, more work needs to be done to more fully understand the manner in which the lower bound $B_0(m,d)$ depends on m and d.

The important first step toward this goal is establishing the following convexity property of the function $k\mapsto \mathrm{bayes}(k,b)$.

PROPOSITION 2.8. *Take any* $b \in [0,1]$. *Then the largest convex function* $[0,\infty) \ni \kappa \mapsto \overset{\smile}{\mathrm{bayes}}(\kappa, b)$ *such that* $\overset{\smile}{\mathrm{bayes}}(k,b) \le \mathrm{bayes}(k,b)$ *for all* $k \in \{0,1,\dots\}$ *is given by the formula*

$$
\begin{aligned}
&\overset{\smile}{\mathrm{bayes}}(\kappa, b) \\
&:= \begin{cases} (1-\kappa)\,\mathrm{bayes}(0,b) + \kappa\,\mathrm{bayes}(1,b) = \dfrac{1}{2}(1-\kappa b) \\ \qquad \text{if } 0 \le \kappa \le 1, \\ \dfrac{2i+3-\kappa}{2}\mathrm{bayes}(2i+1,b) + \dfrac{\kappa-2i-1}{2}\mathrm{bayes}(2i+3,b) \\ \qquad \text{if } 2i+1 \le \kappa \le 2i+3 \end{cases}
\end{aligned}
\tag{2.25}
$$

for any $i \in \overline{0,\infty}$.

That is, the largest convex minorant $\overset{\smile}{\mathrm{bayes}}(\cdot,b)$ on $[0,\infty)$ of the function $\mathrm{bayes}(\cdot,b)$ on $\{0,1,\dots\}$ is just the linear interpolation of $\mathrm{bayes}(\cdot,b)$ at the points $0,1,3,5,\dots$. This is illustrated in Figure 1.

Recall the definition (1.7) of ν. Using (2.21), Proposition 2.8, Jensen's inequality and the equality $\mathsf{E}\,N = \nu$, one immediately obtains the following.

THEOREM 2.9.

$$
\inf_L \sup_D \mathfrak{R}(L,D) \ge B_0(m,d) \ge B_1(\nu) := \sup_{b\in(0,1)} b\,\overset{\smile}{\mathrm{bayes}}(\nu,b). \tag{2.26}
$$

Here and in the rest of this section, $\inf_L$ can be replaced by either $\inf_{L\in\mathcal{L}}$ or $\inf_{L\in\mathcal{L}_{\mathsf{rand}}}$.

REMARK 2.10. An advantage of the lower bound $B_1(\nu)$ in (2.26) over the bound $B_0(m,d)$ in (2.21) is that it depends only on $\nu = m/d$; also, $B_1(\nu)$ is not hard to compute unless ν is too large. Yet, the nature of the dependence of $B_1(\nu)$ on

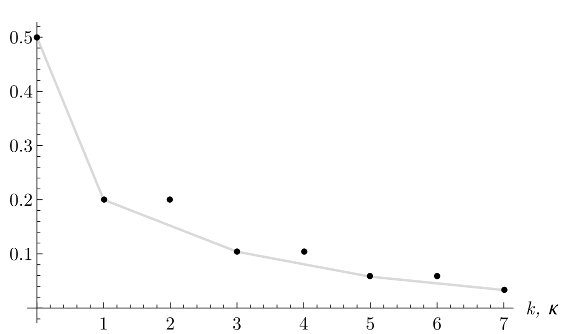

FIG. 1. *Graphs of the maps* $\{0,1,\dots,7\} \ni k \mapsto \mathrm{bayes}(k,b)$ (*black dots*) *and* $[0,7] \ni \kappa \mapsto \overset{\smile}{\mathrm{bayes}}(\kappa,b)$ (*gray broken line*) *for* $b = 0.6$.

ν may still seem rather obscure. Therefore, we are going to present a lower bound on $B_1(\nu)$ that is much easier to grasp and yet is (i) asymptotic to the original lower bound $B_0(m,d)$ for $\nu = m/d \to \infty$ and (ii) close to $B_0(m,d)$ even for moderate values of $\nu = m/d$.

In Appendix A, we shall obtain explicit and rather tight lower bounds on the function bayes. In view of Theorem 2.9 and Proposition 2.8, this will result in explicit lower bounds on the minimax excess risk $\inf_L \sup_D \mathfrak{R}(L, D)$, as follows.

Let

$$z_* = 0.75179\ldots \tag{2.27}$$

be the unique maximizer of $\frac{z}{2}(1 - \operatorname{erf}(z/\sqrt{2}))$ in real $z > 0$, with the maximum value $c_\infty = 0.16997\ldots$, as in (2.1).

THEOREM 2.11. *Assume that $\nu \geq 1$. Let $i_\nu := \lfloor \frac{\nu-1}{2} \rfloor$. Then*

$$\inf_L \sup_D \mathfrak{R}(L, D) \geq B_1(\nu) \geq B_2(\nu) := \frac{c_\nu}{\sqrt{\nu}}, \tag{2.28}$$

where

$$\begin{aligned} c_\nu &:= \frac{z_*}{2}\left(1 - C_{i_\nu}\frac{\operatorname{erf}(z_*/\sqrt{2})}{\exp\{-z_*^2/(6\nu)\}}\right) < c_\infty \quad \textit{and} \\ C_i &= \frac{\sqrt{\pi(i+1/2)}}{2^{2i}}\binom{2i}{i} \end{aligned} \tag{2.29}$$

for $i = 0, 1, \ldots$. Moreover, for $\nu \geq 3$, $B_2(\nu)$ admits a simple lower bound on it:

$$\begin{aligned} B_2(\nu) &\geq \tilde{B}_2(\nu) := \frac{\tilde{c}_\nu}{\sqrt{\nu}}, \qquad \textit{where} \\ \tilde{c}_\nu &:= \frac{z_*}{2}\left(1 - \left(\frac{i_\nu+1}{i_\nu}\right)^{1/8}\frac{\operatorname{erf}(z_*/\sqrt{2})}{\exp\{-z_*^2/(6\nu)\}}\right) \leq c_\nu. \end{aligned} \tag{2.30}$$

REMARK 2.12. To obtain the second inequality in (2.28) ($B_1(\nu) \geq B_2(\nu)$), in the proof of Theorem 2.11 we are going to use, in particular, two facts: (i) that C_i decreases in i (as stated in Lemma A.2) and (ii) the concavity of $\operatorname{erf}(b\sqrt{k/2})$ in k. If one also uses the obvious fact that $\operatorname{erf}(b\sqrt{k/2})$ increases in k, then, by Chebyshev's integral inequality,

$$\begin{aligned} &(1-w_i)C_i \operatorname{erf}\left(b\sqrt{\frac{2i+1}{2}}\right) + w_i C_{i+1} \operatorname{erf}\left(b\sqrt{\frac{2i+3}{2}}\right) \\ &\quad \leq \left[(1-w_i)C_i + w_i C_{i+1}\right] \end{aligned}$$

$$\times \left[(1-w_i)\,\mathrm{erf}\left(b\sqrt{\frac{2i+1}{2}}\right) + w_i\,\mathrm{erf}\left(b\sqrt{\frac{2i+3}{2}}\right)\right]$$
$$\le \left[(1-w_i)C_i + w_i C_{i+1}\right]\mathrm{erf}(b\sqrt{\nu}),$$

where $i := i_\nu$ and $w_i := \frac{\nu-2i-1}{2} \in [0,1)$. Thus, one can replace $C_{i_\nu} = C_{i_\nu} \vee C_{i_\nu+1}$ in (2.29) by the smaller (and hence better) value $(1-w_i)C_i + w_i C_{i+1}$, with $i = i_\nu$. Quite similarly, one can replace $\tilde{C}_{i_\nu} := (\frac{i_\nu+1}{i_\nu})^{1/8} = \tilde{C}_{i_\nu} \vee \tilde{C}_{i_\nu+1}$ in (2.30) by the smaller (and hence better) value $(1-w_i)\tilde{C}_i + w_i\tilde{C}_{i+1}$, with $i = i_\nu$. However, these improvements are comparatively small, especially for larger values of ν, and the resulting expressions will be less easy to perceive.

It is clear that

$$c_\nu \to c_\infty \quad \text{and} \quad \tilde{c}_\nu \to c_\infty \tag{2.31}$$

as $\nu \to \infty$. In fact, c_ν and even $\tilde{c}_\nu$ are rather close to c_∞ already for rather small values of ν. For example, one has $c_5 = 0.15536\ldots$, $\tilde{c}_5 = 0.15514\ldots$, $c_{50} = 0.16852\ldots$ and $\tilde{c}_{50} = 0.16852\ldots$, and indeed all these four values are rather close to $c_\infty = 0.16997\ldots$. We also see that the values of $\tilde{c}_\nu$ are not only simpler to compute than, but also very close to, the corresponding values of c_ν.

Inequality (2.28) in Theorem 2.11 does not cover the case $0 < \nu < 1$, and inequality (2.30) does not cover the case $0 < \nu < 3$. These two apparently less important cases are covered, complementarily, by the following.

PROPOSITION 2.13.

$$\inf_L \sup_D \mathfrak{R}(L,D) \ge \hat{B}_2(\nu) \tag{2.32}$$

$$:= \begin{cases} B_1(\nu) = \dfrac{1}{2}(1-\nu) & \textit{if } 0 < \nu \le \dfrac{1}{2}, \\ B_1(\nu) = \dfrac{1}{8\nu} & \textit{if } \dfrac{1}{2} \le \nu \le 1, \\ \dfrac{(17-2\nu)(57{,}187 - 3253\nu - 138\nu^2 + 212\nu^3 - 8\nu^4)}{6{,}480{,}000} & \\ \qquad\qquad \textit{if } 1 \le \nu \le 3. \end{cases} \tag{2.33}$$

REMARK 2.14. In particular, $\hat{B}_2(1) = B_1(1) = 0.125$, $\hat{B}_2(3) = 0.087018\ldots = \frac{0.15072\ldots}{\sqrt{3}}$, and $B_1(3) = 0.087019\ldots = \frac{0.15072\ldots}{\sqrt{3}}$ (cf. (2.28)). More generally, the choices $b = 1$ for $\nu \in (0, \frac{1}{2}]$ and $b = \frac{1}{2\nu}$ for $\nu \in [\frac{1}{2}, 1]$ in the proof of Proposition 2.13 are optimal, in the sense that $\hat{B}_2(\nu) = B_1(\nu)$ for $\nu \in (0,1]$, as indicated in (2.32). The choice $b = \frac{1}{30}(17-2\nu)$ for $\nu \in [1,3]$ in the just mentioned proof is nearly optimal; namely, then $\hat{B}_2(\nu) > B_1(\nu) - 2\times 10^{-6}$, for all $\nu \in [1,3]$; see details on this remark in Section 3, right after the proof of Proposition 2.13.

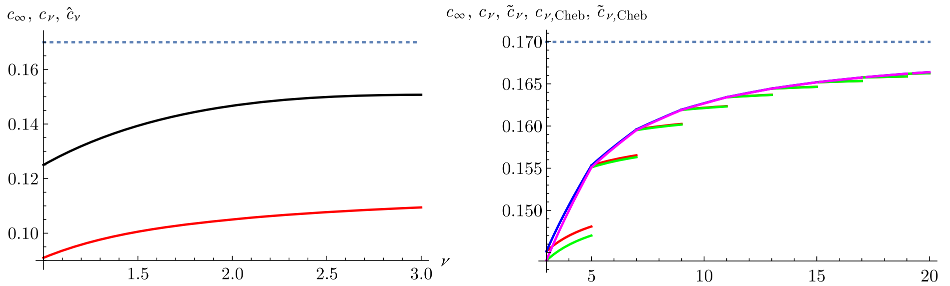

FIG. 2. *Left panel*: *graphs of* c_ν (*red*) *and* $\hat{c}_\nu := \sqrt{\nu}\hat{B}_2(\nu)$ (*black*) *for* $\nu \in [1,3]$. *Right panel*: *graphs of* c_ν (*red*), $\tilde{c}_\nu$ (*green*), $c_{\nu,\text{Cheb}}$ (*blue*) *and* $\tilde{c}_{\nu,\text{Cheb}}$ (*magenta*) *for* $\nu \in [3,20]$, *where* $c_{\nu,\text{Cheb}}$ *and* $\tilde{c}_{\nu,\text{Cheb}}$ *are obtained from the expressions for* c_ν *and* $\tilde{c}_\nu$ *in* (2.29) *and* (2.30) *by replacing there* C_{i_ν} *and* $\tilde{C}_{i_\nu} = (\frac{i_\nu+1}{i_\nu})^{1/8}$ *by the "Chebyshev" expressions* $(1-w_i)C_i + w_iC_{i+1}$ *and* $(1-w_i)\tilde{C}_i + w_i\tilde{C}_{i+1}$, *with* $i = i_\nu$ *and* $w_i := \frac{\nu-2i-1}{2}$, *as discussed in Remark* 2.12. *The dotted horizontal line in both panels is at the level of* $c_\infty = 0.16997\ldots$.

Of course, one can also rather easily give an exact algebraic expression for $B_1(\nu)$ with $\nu \in [1,3]$; however, that expression (in terms of certain roots of certain polynomials in one variable whose coefficients are polynomials in ν) is complicated and, therefore, omitted here.

Theorem 2.11, Remark 2.12, relations (2.31), Proposition 2.13 and Remark 2.14 are illustrated in Figure 2.

Let us also present the following very simple, but suboptimal, lower bound; cf., for example, (2.18).

PROPOSITION 2.15. *If* $\nu \geq \frac{3}{41}$, *then*

$$\inf_L \sup_D \mathfrak{R}(L,D) \geq B_0(m,d) \geq \frac{0.125}{\sqrt{\nu}}. \tag{2.34}$$

Note that the restriction $\nu \geq \frac{3}{41}$ in Proposition 2.15 cannot be dropped, and it is in fact rather close to necessity. Indeed, in view of Remark 2.3, the lower bound $\frac{0.125}{\sqrt{\nu}}$ in (2.34) cannot hold for $\nu < \frac{1}{16} = \frac{3}{48}$. The lower bound $\frac{0.125}{\sqrt{\nu}}$ in (2.34) was obtained by different methods in [2] (under the condition $\nu \geq \frac{1}{4}$, which was omitted from the paper but stated in a corrigendum).

In conclusion of this section, we summarize the asymptotic behavior of the lower bounds $B_0(m,d)$, $B_1(\nu)$, $B_2(\nu)$, $\tilde{B}_2(\nu)$ on the minimax EER, as well as the asymptotic behavior of the minimax EER itself.

THEOREM 2.16.

$$\frac{c_\infty}{\sqrt{\nu}} \sim \inf_L \sup_D \mathfrak{R}(L,D) \geq B_0(m,d) \geq B_1(\nu) \geq B_2(\nu) \geq \tilde{B}_2(\nu) \sim \frac{c_\infty}{\sqrt{\nu}} \tag{2.35}$$

as m *and* d *vary in any way such that* $\nu = m/d \to \infty$.

Thus, in view of (2.2), the limit relation in (2.1) holds and, moreover, all the lower bounds $B_0(m,d)$, $B_1(\nu)$, $B_2(\nu)$, $\tilde{B}_2(\nu)$ on the minimax EER are asymptotically equivalent to the minimax EER itself whenever $\nu = m/d \to \infty$. Clearly, Theorem 2.16 complements Theorem 2.4.

3. Proofs. In this section, we shall prove (or provide details for) Theorems 2.2 and 2.6, Proposition 2.8, Theorem 2.11, Proposition 2.13, Remark 2.14, Proposition 2.15, Theorems 2.4 and 2.16 (together), and finally Theorem 2.1, in this order.

PROOF OF THEOREM 2.2. The first equality in (2.13) can be obtained using the von Neumann minimax duality theorem for bilinear functions on the product of simplexes [23] (plus a certain symmetrization argument); more general minimax duality theorems, for convex-concave-like functions, were given in [21], and in [17] a necessary and sufficient condition for the minimax duality for such functions was given.

However, here we are going to offer a more direct and explicit argument, using the explicit form of the to-be-proved-minimax decision rule L^*_{ERM}, as defined in (2.11).

To gain some insight, let us begin with the simple case $d = 1$. In that case, $\mathcal{X} = [d] = [1] = \{1\}$, $p = (1)$ (that is, $p_1 = 1$) and $\beta = (b)$ with $b := \beta_1 \in [-1, 1]$; also, in the just mentioned definition (2.11) of L^*_{ERM}, the terms x, v_x, i_x, and n_x simplify, respectively, to 1, $v_1 = \sum_{i=1}^m y_i$, 1, and $n_1 = m$, in accordance with the definitions of v_x, i_x and n_x in (2.8) and (2.12). Here, we also have $X_i = 1$ for all i and hence, in view of (2.4) and (2.3), $Z_m^{p,\beta}$ equals

$$Z_m^b := \big((1, Y_1^b), \dots, (1, Y_m^b)\big) \tag{3.1}$$

in distribution, where the Y_i^b's are as in the statement of Theorem 2.2.

A standard argument (see, e.g., [7], Section 1.8) shows that $L^*_{1,\mathsf{ERM}}$ is an optimal Bayesian decision rule, in the sense of being a minimizer of the average $\frac{1}{2}\sum_{y\in\mathcal{Y}} \mathsf{P}(L(Z_m^{|b|y}, U)(1) \neq y)$ of the types I and II error probabilities over all $L \in \mathcal{L}_{\mathsf{rand},1}$, that is, over all randomized learning algorithms L for $d = 1$. By symmetry, without loss of generality $b \geq 0$, and so, $b \in [0, 1]$. Then for the corresponding Bayes risk, given by the expression $\frac{1}{2}\sum_{y\in\mathcal{Y}} \mathsf{P}(L^*_{1,\mathsf{ERM}}(Z_m^{by}, U)(1) \neq y)$, for $m \geq 1$ one has

$$\begin{aligned}
&2 \cdot \frac{1}{2} \sum_{y\in\mathcal{Y}} \mathsf{P}\big(L\big(Z_m^{|b|y}, U\big)(1) \neq y\big) \\
&\quad\geq 2 \cdot \frac{1}{2} \sum_{y\in\mathcal{Y}} \mathsf{P}\big(L^*_{1,\mathsf{ERM}}\big(Z_m^{by}, U\big)(1) \neq y\big) \\
&\quad= \mathsf{P}\big(V_m^b < 0\big) + \mathsf{P}\big(V_m^b = 0, Y_1^b < 0\big) \\
&\qquad+ \mathsf{P}\big(V_m^{-b} > 0\big) + \mathsf{P}\big(V_m^{-b} = 0, Y_1^{-b} > 0\big)
\end{aligned}$$

$$
\begin{aligned}
&= \mathsf{P}\big(V_m^b < 0\big) + \mathsf{P}\big(V_m^b = 0\big) + \mathsf{P}\big(V_m^{-b} > 0\big) \\
&= \mathsf{P}\big(V_m^b \le 0\big) + \mathsf{P}\big(V_m^{-b} > 0\big) \\
&= 1 - \big(\mathsf{P}\big(V_m^b > 0\big) - \mathsf{P}\big(V_m^{-b} > 0\big)\big) = 2\,\mathrm{bayes}(m,b),
\end{aligned} \tag{3.2}
$$

in accordance with (2.16) (implying, in particular, that (Y_1^{-b}, V_m^{-b}) equals $(-Y_1^b, -V_m^b)$ in distribution), (2.14), (2.15) and the assumption $b \in [0,1]$ (which implies $\mathsf{P}(V_m^b > 0) \ge \mathsf{P}(V_m^{-b} > 0)$, since Y_i^b is stochastically increasing in b). Thus, the Bayes risk $\frac{1}{2}\sum_{y\in\mathcal{Y}} \mathsf{P}(L^*_{1,\mathsf{ERM}}(Z_m^{by}, U)(1) \ne y)$ equals $\mathrm{bayes}(m,b)$ for $b \in [0,1]$ and $m \ge 1$. This conclusion also trivially holds for $m = 0$ (in which case $\mathrm{bayes}(m,b) = \frac{1}{2}$).

Moreover, for each $m \in \overline{0,\infty}$, the Bayes rule $L^*_{1,\mathsf{ERM}}$ is a risk equalizer, in the sense that

$$
\begin{aligned}
&\text{for } b \in [0,1] \text{ and } y \in \{-1,1\}, \\
&\mathsf{P}\big(L^*_{1,\mathsf{ERM}}\big(Z_m^{by}, U\big)(1) \ne y\big) = \mathrm{bayes}(m,b),
\end{aligned} \tag{3.3}
$$

which does not depend on the choice of y; this conclusion follows because (i) $(L^*_{1,\mathsf{ERM}}((Z_m^b)^-, -U) = -L^*_{1,\mathsf{ERM}}(Z_m^b, U)$, where $(Z_m^b)^- := ((1, -Y_1^b), \dots, (1, -Y_m^b))$, and (ii) the distribution of $(Y_1^{-b}, \dots, Y_k^{-b}, -U)$ is the same as that of $-(Y_1^b, \dots, Y_k^b, U)$.

Let us now proceed to the general case of any natural d, which in a sense reduces to the case $d = 1$. Take any $m \in \overline{0,\infty}$, any randomized learning algorithm $L\colon (\mathcal{X} \times \mathcal{Y})^m \times [-1,1] \to \mathcal{H}$, any $p \in [0,1]^{[d]}$ such that $\sum_{x=1}^d p_x = 1$, and any $\beta \in [-1,1]^{[d]}$. For each $x \in [d]$, introduce the random set

$$
\mathcal{J}_x^p := \{i \in [m]\colon X_i^p = x\} \tag{3.4}
$$

and its cardinality

$$
N_x^p := \mathrm{card}\, \mathcal{J}_x^p. \tag{3.5}
$$

Then, by (2.10),

$$
\mathfrak{R}_m(L; p, \beta) = \sum_{x=1}^d p_x |\beta_x| \sum_{k=0}^m \mathsf{P}\big(L\big(Z_m^{p,\beta}, U\big)(x) \ne h_\beta(x), N_x^p = k\big). \tag{3.6}
$$

Next, take any $x \in [d]$ and any $k = 0, \dots, m$. Then

$$
\begin{aligned}
&\mathsf{P}\big(L\big(Z_m^{p,\beta}, U\big)(x) \ne h_\beta(x), N_x^p = k\big) \\
&\quad = \sum_{J \in \binom{[m]}{k}} \mathsf{P}\big(L\big(Z_m^{p,\beta}, U\big)(x) \ne h_\beta(x), \mathcal{J}_x^p = J\big),
\end{aligned} \tag{3.7}
$$

where $\binom{[m]}{k} := \{J \subseteq [m]\colon \mathrm{card}\, J = k\}$.

Further, take any set $J \in \binom{[m]}{k}$. Writing J as $\{i_1, \dots, i_k\}$ with $i_1 < \cdots < i_k$, let $X_J^p := (X_{i_1}^p, \dots, X_{i_k}^p)$, and similarly define $X_{J^c}^p$, $Y_J^{p,\beta}$, and $Y_{J^c}^{p,\beta}$, where $J^c := [m] \setminus J$. Let also $Z_{J^c}^{p,\beta} := (X_{J^c}^p, Y_{J^c}^{p,\beta})$. For any $x \in \mathcal{X}$, let $x^J := (x, \dots, x) \in \mathcal{X}^k$ and $\mathcal{Z}_x := (\mathcal{X} \setminus \{x\}) \times \mathcal{Y}$. Then, in view of (2.5),

$$\begin{aligned} &\mathsf{P}\big(L\big(Z_m^{p,\beta}, U\big)(x) \neq h_\beta(x), \mathcal{J}_x^p = J\big) \\ &\quad = \sum_{z \in \mathcal{Z}_x^{m-k}} \mathsf{P}\big(L\big(Z_m^{p,\beta}, U\big)(x) \neq \operatorname{sgn} \beta_x | X_J^p = x^J, Z_{J^c}^{p,\beta} = z\big) \\ &\qquad \times \mathsf{P}\big(X_J^p = x^J, Z_{J^c}^{p,\beta} = z\big). \end{aligned} \tag{3.8}$$

(Here and in what follows, to simplify the writing, we neglect the possibility that $\mathsf{P}(X_J^p = x^J, Z_{J^c}^{p,\beta} = z)$ may equal 0. Of course, in such cases we may let the corresponding conditional probabilities in (3.8) take whatever values deemed most suitable for us at any given point.)

By (3.6), (3.7) and (3.8),

$$\begin{aligned} &\mathfrak{R}_m(L; p, \beta) \\ &\quad = \sum_{x,k,J,z} p_x |\beta_x| \mathsf{P}\big(L\big(Z_m^{p,\beta}, U\big)(x) \neq \operatorname{sgn} \beta_x | X_J^p = x^J, Z_{J^c}^{p,\beta} = z\big) \\ &\qquad \times \mathsf{P}\big(X_J^p = x^J, Z_{J^c}^{p,\beta} = z\big), \end{aligned} \tag{3.9}$$

where $\sum_{x,k,J,z} := \sum_{x=1}^d \sum_{k=0}^m \sum_{J \in \binom{[m]}{k}} \sum_{z \in \mathcal{Z}_x^{m-k}}$ and, moreover, the sum

$$\sum_{J,z} \mathsf{P}\big(X_J^p = x^J, Z_{J^c}^{p,\beta} = z\big) = \mathsf{P}\big(N_x^p = k\big) \tag{3.10}$$

(where $\sum_{J,z} := \sum_{J \in \binom{[m]}{k}} \sum_{z \in \mathcal{Z}_x^{m-k}}$) does not depend on β. To quickly see why identity (3.10) holds, look back at (3.7) and (3.8), with the event $\{L(Z_m^{p,\beta}, U)(x) \neq h_\beta(x)\}$ replaced there by an event of probability 1.

Since $(X_1^p, Y_1^{p,\beta}), \dots, (X_m^p, Y_m^{p,\beta}), U$ are independent, for any $L \in \mathcal{L}_{\mathsf{rand}}$, $x \in [d]$, $k \in \overline{0,m}$, $J \in \binom{[m]}{k}$, $z \in \mathcal{Z}_x^{m-k}$, and $p \in [0,1]^{[d]}$ such that $\sum_{x=1}^d p_x = 1$, the conditional probability $\mathsf{P}(L(Z_m^{p,\beta}, U)(x) \neq \operatorname{sgn} \beta_x | X_J^p = x^J, Z_{J^c}^{p,\beta} = z)$ in (3.8) depends on β only through β_x, whereas the unconditional probability $\mathsf{P}(X_J^p = x^J, Z_{J^c}^{p,\beta} = z)$ in (3.8) depends on β only through $\beta_{\setminus x} := \beta|_{\mathcal{X} \setminus \{x\}}$—the restriction of the function β to the subset $\mathcal{X} \setminus \{x\}$ of the set $\mathcal{X}$. So, introducing the averaging operators

$$\operatorname*{ave}_\sigma := \frac{1}{2^d} \sum_{\sigma \in \{-1,1\}^{[d]}},$$

$$\operatorname*{ave}_{\sigma_{\setminus x}} := \frac{1}{2^{d-1}} \sum_{\sigma_{\setminus x} \in \{-1,1\}^{\mathcal{X} \setminus \{x\}}},$$

$$\operatorname*{ave}_{\sigma_x} := \frac{1}{2} \sum_{\sigma_x \in \{-1,1\}},$$

in view of (3.9) one has

$$\begin{aligned}
&\operatorname*{ave}_{\sigma} \mathfrak{R}_m\big(L; p, |\beta|\sigma\big) \\
&\quad = \sum_{x,k,J,z} p_x |\beta_x| \operatorname*{ave}_{\sigma_x} \mathsf{P}\big(L\big(Z_m^{p,|\beta|\sigma}, U\big)(x) \neq \sigma_x | X_J^p = x^J, Z_{J^{\mathsf{c}}}^{p,|\beta|\sigma} = z\big) \\
&\qquad \times \operatorname*{ave}_{\sigma_{\setminus x}} \mathsf{P}\big(X_J^p = x^J, Z_{J^{\mathsf{c}}}^{p,|\beta|\sigma} = z\big).
\end{aligned} \tag{3.11}$$

Recall that the random pairs $(X_1^p, Y_1^{p,\beta}), \dots, (X_m^p, Y_m^{p,\beta})$ are independent copies of the random pair $(X^p, Y^{p,\beta}) = (X^D, Y^D)$ satisfying condition (2.3), and the r.v. U is independent of these pairs. So, for any $x \in [d]$, $k \in \overline{0,m}$, $J \in \binom{[m]}{k}$, $z \in \mathcal{Z}_x^{m-k}$ and $p \in [0,1]^{[d]}$ such that $\sum_{x=1}^d p_x = 1$, the conditional distribution of $(Y_J^{p,\beta}, U)$ given $X_J^p = x^J$ and $Z_{J^{\mathsf{c}}}^{p,\beta} = z$ is the same as the distribution of $(Y_1^{\beta_x}, \dots, Y_k^{\beta_x}, U)$, where the Y_i^b's are again as in the statement of Theorem 2.2. Therefore, in view of (3.2), for each $x \in [d]$, each $z = ((x_{k+1}, y_{k+1}), \dots, (x_m, y_m)) \in \mathcal{Z}_x^{m-k}$, and $J = [k]$,

$$\begin{aligned}
&\operatorname*{ave}_{\sigma_x} \mathsf{P}\big(L\big(Z_k^{p,|\beta|\sigma}, U\big)(x) \neq \sigma_x | X_J^p = x^J, Z_{J^{\mathsf{c}}}^{p,|\beta|\sigma} = z\big) \\
&\quad = \operatorname*{ave}_{\sigma_x} \mathsf{P}\big(L_x\big(Z_k^{|\beta_x|\sigma_x}, U\big)(1) \neq \sigma_x\big) \geq \operatorname{bayes}\big(k, |\beta_x|\big),
\end{aligned} \tag{3.12}$$

where Z_k^b is as in (3.1) and $L_x\colon ([1] \times \mathcal{Y})^k \times [-1,1] \to \{-1,1\}^{[1]}$ is the randomized learning algorithm (for the case $d = 1$) defined by the formula $L_x(((1, y_1), \dots, (1, y_k)), u) := h|_{\{x\}}$, the restriction of the function h to the singleton set $\{x\}$, where $h := h_{w;u} := L(w, u) \in \{-1,1\}^{[d]}$, $w := ((x, y_1), \dots, (x, y_k), (x_{k+1}, y_{k+1}), \dots, (x_m, y_m))$ and $u \in [-1,1]$. Clearly then, (3.12) holds for any $J \in \binom{[m]}{k}$.

Similarly, but now using (3.3) instead of (3.2), we have

$$\mathsf{P}\big(L_{m,\mathsf{ERM}}^*\big(Z_m^{p,|\beta|\sigma}, U\big)(x) \neq \sigma_x | X_J^p = x^J, Z_{J^{\mathsf{c}}}^{p,|\beta|\sigma} = z\big) = \operatorname{bayes}\big(k, |\beta_x|\big) \tag{3.13}$$

for each $x \in [d]$ *and each* $\sigma_x \in \{-1,1\}$, and hence the average of the left-hand side (LHS) of (3.13) over $\sigma_x \in \{-1,1\}$ equals $\operatorname{bayes}(k, |\beta_x|)$ as well. Note also that $\mathsf{P}(X_J^p = x^J, Z_{J^{\mathsf{c}}}^{p,|\beta|\sigma} = z)$ does not depend on L. So, collecting identity (3.11), its counterpart with $L_{m,\mathsf{ERM}}^*$ in place of L, (3.12), and identity (3.13) with its LHS replaced by the average of that LHS over $\sigma_x \in \{-1,1\}$, we see that

$$\operatorname*{ave}_{\sigma} \mathfrak{R}_m\big(L; p, |\beta|\sigma\big) \geq \operatorname*{ave}_{\sigma} \mathfrak{R}_m\big(L_{m,\mathsf{ERM}}^*; p, |\beta|\sigma\big). \tag{3.14}$$

Moreover, by (3.9) with $L^*_{m,\mathsf{ERM}}$ in place of L, (3.13) with $\sigma = \operatorname{sgn}\beta$, and (3.10),

$$\begin{aligned}
\mathfrak{R}_m(L^*_{m,\mathsf{ERM}}; p, \beta) &= \sum_{x,k,J,z} p_x|\beta_x| \operatorname{bayes}(k, |\beta_x|) \mathsf{P}(X^p_J = x^J, Z^{p,\beta}_{J^c} = z) \\
&= \sum_{x=1}^{d} p_x|\beta_x| \sum_{k=0}^{m} \operatorname{bayes}(k, |\beta_x|) \sum_{J,z} \mathsf{P}(X^p_J = x^J, Z^{p,\beta}_{J^c} = z) \\
&= \sum_{x=1}^{d} p_x|\beta_x| \sum_{k=0}^{m} \operatorname{bayes}(k, |\beta_x|) \mathsf{P}(N^p_x = k) \\
&= \sum_{x=1}^{d} p_x|\beta_x| \mathsf{E} \operatorname{bayes}(N^p_x, |\beta_x|), \qquad (3.15)
\end{aligned}$$

which proves (2.17) and the second equality in (2.13) (here one may recall the definition of $\mathfrak{R}_m(L; p, \beta)$ in (2.10)). We also see that $\mathfrak{R}_m(L^*_{m,\mathsf{ERM}}; p, \beta)$ depends on β only through $|\beta|$. This and (3.14) yield

$$\begin{aligned}
&\max_{\sigma\in\{-1,1\}^d} \mathfrak{R}_m(L^*_{m,\mathsf{ERM}}; p, |\beta|\sigma) \\
&\quad= \operatorname*{ave}_{\sigma} \mathfrak{R}_m(L^*_{m,\mathsf{ERM}}; p, |\beta|\sigma) \\
&\quad= \sum_{x=1}^{d} p_x|\beta_x| \mathsf{E} \operatorname{bayes}(N^p_x, |\beta_x|) \\
&\quad\le \operatorname*{ave}_{\sigma} \mathfrak{R}_m(L; p, |\beta|\sigma) \le \max_{\sigma\in\{-1,1\}^d} \mathfrak{R}_m(L; p, |\beta|\sigma).
\end{aligned}$$

Taking now $\sup_{p,\beta}$, we see that $\sup_D \mathfrak{R}_m(L^*_{m,\mathsf{ERM}}, D) \le \sup_D \mathfrak{R}_m(L, D)$ for all L, which proves the first equality in (2.13). This completes the proof of Theorem 2.2. $\square$

PROOF OF THEOREM 2.6. This theorem follows immediately from (3.14) and (3.15) by taking there $p_x = \frac{1}{d}$ for all $x \in [d]$ and any $\beta \in \{-b, b\}^{[d]}$. $\square$

PROOF OF PROPOSITION 2.8. By (2.14)–(2.15), $\operatorname{bayes}(k, b) = 1/2$ for all $k = 0, 1, \dots$ if $b = 0$. Suppose now that $b \in (0, 1]$. Letting $k \to \infty$ and using the law of large numbers, in view of (2.16) we have $\frac{1}{k} V^b_k \to \mathsf{E}\, Y^b_1 = b > 0$ in probability. So, $\mathsf{P}(V^b_k > 0) \to 1$; similarly, $\mathsf{P}(V^{-b}_k > 0) \to 0$. Recalling (2.14)–(2.15) again, we see that $\operatorname{bayes}(k, b) \to 0$. So, by Lemma A.1 in Appendix A, $\operatorname{bayes}(k, b)$ is convex and nonincreasing in $k \in \{0, 1, 3, 5, \dots\}$, for each $b \in [0, 1]$. It remains to use relations (A.9) and (A.7) in Appendix A and, again, (2.14). $\square$

PROOF OF THEOREM 2.11. The first inequality in (2.28) comes from (2.26). The second inequality in (2.28) follows immediately from Lemma A.3 with $\kappa = \nu$

and $b = z_*/\sqrt{\nu}$. The inequality in (2.29) holds because, by Lemma A.2, $C_i > 1$ and, by the paragraph containing (2.27), $c_\infty = \frac{z_*}{2}(1 - \operatorname{erf}(z_*/\sqrt{2}))$. The last equality in (2.29) follows by (A.12). The inequalities in (2.30) follow immediately from (2.28), (2.29) and (A.13). $\square$

PROOF OF PROPOSITION 2.13. By (2.25), (2.14) and (A.9),

$$\widetilde{\text{bayes}}(\nu) = \begin{cases} \frac{1}{2}(1 - \nu b) & \text{if } 0 < \nu \le 1, \\ \frac{1}{8}\big(4 + b^3(\nu - 1) - b(3 + \nu)\big) & \text{if } 1 \le \nu \le 3. \end{cases}$$

Recalling now (2.26) and using the values $b = 1$ for $\nu \in (0, \frac{1}{2}]$, $b = \frac{1}{2\nu}$ for $\nu \in [\frac{1}{2}, 1]$, and and $b = \frac{1}{30}(17 - 2\nu)$ for $\nu \in [1, 3]$, one obtains (2.32). $\square$

DETAILS ON REMARK 2.14. The inequality $\hat{B}_2(\nu) > B_1(\nu) - 2 \times 10^{-6}$ for all $\nu \in [1, 3]$, mentioned in that remark, can be verified, for example, by issuing the Mathematica command `Reduce[b 1/8 (4 + b^3 (nu - 1) - b (3 + nu)) - hB2[nu] >= 2 10^(-6) && 1 <= nu <= 3 && 0 <= b <= 1]`, where `hB2[nu]` stands for $\hat{B}_2(\nu)$. This command then outputs `False`, which means that indeed $B_1(\nu) = \max_{0 \le b \le 1} b\widetilde{\text{bayes}}(\nu) = \max_{0 \le b \le 1} b\frac{1}{8} \times (4 + b^3(\nu - 1) - b(3 + \nu)) < \hat{B}_2(\nu) + 2 \times 10^{-6}$. $\square$

PROOF OF PROPOSITION 2.15. The first inequality in (2.34) holds by (2.21). Take now any $b \in (0, 1]$. From (A.5), Lemma A.2 and inequality $S_q(b) \le b$, it follows that

$$s_k(b) \le b\sqrt{k} \tag{3.16}$$

for odd natural k. By (A.7), inequality (3.16) holds for even natural k as well, and it trivially holds for $k = 0$. Using now the definition of $B_0(m, d)$ in (2.21) together with (2.14), (3.16) and Jensen's inequality, noticing that $\frac{1}{2\sqrt{\nu}} \in (0, 1]$ if $\nu \ge \frac{1}{4}$, and substituting $\frac{1}{2\sqrt{\nu}}$ for b, one has

$$\begin{aligned} B_0(m, d) &\ge b\,\mathsf{E}\,\text{bayes}(N, b) \ge \frac{b}{2}(1 - b\,\mathsf{E}\sqrt{N}) \ge \frac{b}{2}(1 - b\sqrt{\mathsf{E}\,N}) \\ &= \frac{b}{2}(1 - b\sqrt{\nu}) = \frac{0.125}{\sqrt{\nu}}, \end{aligned}$$

in the case when $\nu \ge \frac{1}{4}$.

It remains to consider the case when $\frac{1}{4} > \nu \ge \frac{3}{41}$. Then, by (2.26) and (2.32),

$$B_0(m, d) \ge B_1(\nu) = \frac{1}{2}(1 - \nu) > \frac{0.125}{\sqrt{\nu}},$$

which completes the proof of Proposition 2.15. $\square$

PROOF OF THEOREMS 2.4 AND 2.16. The two inequalities in (2.18) are trivial. The first, second, third and fourth inequalities in (2.35) were already established as the inequalities in (2.21), the second inequality in (2.26), the second inequality in (2.28) and the first inequality in (2.30), respectively. The second asymptotic equivalence in (2.35) follows immediately from (2.30) and (2.31).

So, in view of (2.13), it suffices to show that (2.19) holds and

$$B(m,d) \overset{?}{\lesssim} \frac{c_\infty}{\sqrt{\nu}}, \tag{3.17}$$

where, as usual, $A \lesssim B$ means $\limsup A/B \le 1$.

In this proof, all the limit relations are stated for $\nu = m/d \to \infty$ and all the other relations are stated under the condition that ν is large enough.

By (2.10), (2.7), (2.20) and (2.5),

$$\begin{aligned} &\mathfrak{R}(L_{\mathsf{ERM}}; p, \beta) \\ &\quad \le \sum_{x=1}^{d} p_x |\beta_x| \big[\mathsf{P}\big(V_x^{p,\beta} \ne 0, \operatorname{sgn} V_x^{p,\beta} \ne \operatorname{sgn} \beta_x\big) + \mathsf{P}\big(V_x^{p,\beta} = 0\big)\big]. \end{aligned} \tag{3.18}$$

Take now any $b \in [0,1]$ and $k \in \overline{0,\infty}$. In view of (2.16), V_k^b equals $-V_k^{-b}$ in distribution, and V_k^b is stochastically greater than V_k^{-b} (since $b \ge 0$). So, by the definitions (2.14)–(2.15) of $\operatorname{bayes}(k,b)$,

$$\operatorname{bayes}(k,b) = Q_-(k,b) + \frac{1}{2} Q_0(k,b), \tag{3.19}$$

where

$$Q_-(k,b) := \mathsf{P}\big(V_k^b < 0\big) \quad \text{and} \quad Q_0(k,b) := \mathsf{P}\big(V_k^b = 0\big). \tag{3.20}$$

Recalling the definition of the random pairs $(X_1^p, Y_1^{p,\beta}), \dots, (X_m^p, Y_m^{p,\beta})$ in the paragraph containing (2.3) and (2.4), the definition of the random set $\mathcal{J}_x^p$ in (3.4) and the definition of the Y_i^b's in Theorem 2.2, we see that, for any $J \in \binom{[m]}{k}$, the conditional distribution of $(Y_i^{p,\beta})_{i \in J}$ given the event $\{\mathcal{J}_x^p = J\}[= \{X_i = x\ \forall i \in J\} \cup \{X_i \ne x\ \forall i \in J^{\mathsf{c}}\}]$ is the same as the distribution of $(Y_1^{\beta_x}, \dots, Y_k^{\beta_x})$.

Therefore, if $\beta_x > 0$ for some $x \in [d]$, then, in view of (3.5), (2.20), and (3.19),

$$\begin{aligned} &\mathsf{P}\big(V_x^{p,\beta} \ne 0, \operatorname{sgn} V_x^{p,\beta} \ne \operatorname{sgn} \beta_x\big) \\ &\quad = \mathsf{P}\big(V_x^{p,\beta} < 0\big) \\ &\quad = \sum_{k=0}^{m} \mathsf{P}\big(V_x^{p,\beta} < 0, N_x^p = k\big) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=0}^{m} \sum_{J \in \binom{[m]}{k}} \mathsf{P}\big(V_x^{p,\beta} < 0, \mathcal{J}_x^p = J\big) \\
&= \sum_{k=0}^{m} \sum_{J \in \binom{[m]}{k}} \mathsf{P}\bigg(\sum_{i \in J} Y_i^{p,\beta} < 0, \mathcal{J}_x^p = J\bigg) \\
&= \sum_{k=0}^{m} \sum_{J \in \binom{[m]}{k}} \mathsf{P}\bigg(\sum_{i=1}^{k} Y_i^{\beta_x} < 0\bigg) \mathsf{P}\big(\mathcal{J}_x^p = J\big) \\
&= \sum_{k=0}^{m} \mathsf{P}\bigg(\sum_{i=1}^{k} Y_i^{\beta_x} < 0\bigg) \mathsf{P}\big(N_x^p = k\big) = \sum_{k=0}^{m} Q_-\big(k, |\beta_x|\big) \mathsf{P}\big(N_x^p = k\big) \\
&= \mathsf{E}\, Q_-\big(N_x^p, |\beta_x|\big) = \mathsf{E}\, \mathrm{bayes}\big(N_x^p, |\beta_x|\big) - \frac{1}{2} \mathsf{E}\, Q_0\big(N_x^p, |\beta_x|\big).
\end{aligned}
$$

Similarly, the latter expression, $\mathsf{E}\,\mathrm{bayes}(N_x^p, |\beta_x|) - \frac{1}{2}\mathsf{E}\, Q_0(N_x^p, |\beta_x|)$, for $\mathsf{P}(V_x^{p,\beta} \neq 0, \operatorname{sgn} V_x^{p,\beta} \neq \operatorname{sgn} \beta_x)$ holds when $\beta_x < 0$ as well. On the other hand, it is similarly seen that

$$
\mathsf{E}\, Q_0\big(N_x^p, |\beta_x|\big) = \mathsf{P}\big(V_x^{p,\beta} = 0\big).
$$

Thus, (3.18) yields

$$
\mathfrak{R}(L_{\mathsf{ERM}}; p, \beta) \le \sum_{x=1}^{d} p_x |\beta_x|\, \mathsf{E}\, \mathrm{bayes}\big(N_x^p, |\beta_x|\big) + \frac{1}{2} \sum_{x=1}^{d} p_x |\beta_x|\, \mathsf{E}\, Q_0\big(N_x^p, |\beta_x|\big).
$$

In particular, in view of (2.13), this implies the second inequality in (2.19); the first inequality there is trivial.

Now, to complete the proof of (2.19) and Theorem 2.16, it remains to verify (3.17) and

$$
\sup_{p,\beta} \sum_{x=1}^{d} p_x |\beta_x|\, \mathsf{E}\, Q_0\big(N_x^p, |\beta_x|\big) \overset{?}{=} O(1/\nu). \tag{3.21}
$$

Take any $b \in (0, 1]$ and any natural $k \ge 3$, so that, by (A.3) and (A.2), $q := q_k \ge 1 > 0$. Note that $s_k(1) = 1$. Hence, by (2.14) and (A.5),

$$
\begin{aligned}
\mathrm{bayes}(k, b) &= \frac{1}{2}\big(1 - s_k(b)\big) = \frac{1}{2}\big(s_k(1) - s_k(b)\big) \\
&= \frac{1}{2} s_k'(0)\big(S_q(1) - S_q(b)\big) \\
&= \frac{1}{2} s_k'(0) \int_b^1 (1 - u^2)^q \, du \le \frac{1}{2} s_k'(0) \int_b^1 e^{-qu^2} \, du \\
&\le \frac{1}{2} s_k'(0) \int_b^\infty e^{-qu^2} \, du = A_k\big(1 - \mathrm{erf}(b\sqrt{q})\big),
\end{aligned} \tag{3.22}
$$

where $A_k := \frac{s_k'(0)\sqrt{\pi}}{4\sqrt{q}} \to \frac{1}{2}$ as $k \to \infty$, by Lemma A.2 and (A.6). Therefore, for $z := b\sqrt{2q}$ one has

$$b\,\mathrm{bayes}(k,b) \le \frac{\lambda_k}{\sqrt{k}}\frac{z}{2}\big(1-\mathrm{erf}(z/\sqrt{2})\big) \le c_\infty\frac{\lambda_k}{\sqrt{k}} \tag{3.23}$$

by (2.1), where

$$\lambda_k := 2A_k\sqrt{\frac{k}{2q}} \to 1 \tag{3.24}$$

as $k \to \infty$.

Take now any $\varepsilon \in (0,1)$. In view of (3.23) and the first part of Remark 2.3,

$$b\,\mathrm{bayes}(k,b) \le \frac{A}{\sqrt{k+1}} \tag{3.25}$$

for some real $A > 0$, all $b \in [0,1]$, and all $k = 0, 1, \dots$.

Since the r.v. N_x^p has the binomial distribution with parameters m and p_x, one has the following well-known inequality:

$$\mathsf{P}\big(N_x^p \le (1-\varepsilon)mp_x\big) \le e^{-\varepsilon^2 mp_x/2};$$

see, for example, [4], Exercise 2.9; also, this inequality immediately follows from the more general and precise results in [16], (1.3) or (2.31), or [18], Theorem 7.

Therefore,

$$\begin{aligned} S_{01} &:= \sum_{x=1}^d p_x\,\mathsf{E}\frac{1}{\sqrt{N_x^p+1}}\mathrm{I}\{N_x^p \le (1-\varepsilon)mp_x\} \\ &\le \sum_{x=1}^d p_x\,\mathsf{P}\big(N_x^p \le (1-\varepsilon)mp_x\big) \le \sum_{x=1}^d p_x e^{-\varepsilon^2 mp_x/2} \\ &\le \sum_{x=1}^d a_\varepsilon\frac{1}{m} = a_\varepsilon\frac{d}{m}, \end{aligned} \tag{3.26}$$

where $a_\varepsilon := \frac{2}{\varepsilon^2}\sup_{u>0} ue^{-u} = \frac{2}{e\varepsilon^2}$, which depends only on ε. Also,

$$\begin{aligned} S_{02} &:= \sum_{x=1}^d p_x\,\mathsf{E}\frac{1}{\sqrt{N_x^p+1}}\mathrm{I}\{N_x^p > (1-\varepsilon)mp_x\} \\ &\le \sum_{x=1}^d p_x\frac{1}{\sqrt{(1-\varepsilon)mp_x+1}} \le \sum_{x=1}^d \frac{\sqrt{p_x}}{\sqrt{(1-\varepsilon)m}} \le \frac{1}{\sqrt{1-\varepsilon}}\sqrt{\frac{d}{m}}; \end{aligned} \tag{3.27}$$

the last inequality here is obtained using the concavity of the square root function together with the condition $\sum_{x=1}^{d} p_x = 1$. Thus, by (3.26) and (3.27),

$$\sum_{x=1}^{d} p_x \,\mathsf{E}\, \frac{1}{\sqrt{N_x^p + 1}} = S_{01} + S_{02} \le \frac{1}{1-\varepsilon}\sqrt{\frac{d}{m}} \tag{3.28}$$

(if m/d is large enough, depending on ε; recall the framed convention on page 2844).

In view of (3.25),

$$\begin{aligned}
S_{11} &:= \sum_{x=1}^{d} p_x |\beta_x| \,\mathsf{E}\,\mathrm{bayes}\big(N_x^p, |\beta_x|\big)\,\mathrm{I}\big\{N_x^p > (1-\varepsilon)m p_x\big\}\,\mathrm{I}\{p_x \le \varepsilon/d\} \\
&\le \sum_{x=1}^{d} p_x \,\mathsf{E}\, \frac{A}{\sqrt{N_x^p + 1}}\,\mathrm{I}\big\{N_x^p > (1-\varepsilon)m p_x\big\}\,\mathrm{I}\{p_x \le \varepsilon/d\} \\
&\le \sum_{x=1}^{d} p_x \frac{A}{\sqrt{(1-\varepsilon)m p_x + 1}}\,\mathrm{I}\{p_x \le \varepsilon/d\} \le A \sum_{x=1}^{d} \frac{\sqrt{p_x}}{\sqrt{(1-\varepsilon)m}}\,\mathrm{I}\{p_x \le \varepsilon/d\} \\
&\le A \sum_{x=1}^{d} \frac{\sqrt{\varepsilon/d}}{\sqrt{(1-\varepsilon)m}} = A\sqrt{\frac{\varepsilon}{1-\varepsilon}}\sqrt{\frac{d}{m}}.
\end{aligned}$$

Next, taking into account (3.23), (3.24) and (3.27), one has

$$\begin{aligned}
S_{12} &:= \sum_{x=1}^{d} p_x |\beta_x| \,\mathsf{E}\,\mathrm{bayes}\big(N_x^p, |\beta_x|\big)\,\mathrm{I}\big\{N_x^p > (1-\varepsilon)m p_x\big\}\,\mathrm{I}\{p_x > \varepsilon/d\} \\
&\le \sum_{x=1}^{d} p_x |\beta_x| \,\mathsf{E}\,\mathrm{bayes}\big(N_x^p, |\beta_x|\big)\,\mathrm{I}\big\{N_x^p > (1-\varepsilon)m p_x\big\}\,\mathrm{I}\big\{N_x^p > (1-\varepsilon)\varepsilon m/d\big\} \\
&\le \sum_{x=1}^{d} p_x (1+\varepsilon) c_\infty \,\mathsf{E}\, \frac{1}{\sqrt{N_x^p + 1}}\,\mathrm{I}\big\{N_x^p > (1-\varepsilon)m p_x\big\} = (1+\varepsilon) c_\infty S_{02} \\
&\le \frac{(1+\varepsilon)c_\infty}{\sqrt{1-\varepsilon}}\sqrt{\frac{d}{m}}.
\end{aligned}$$

So,

$$\begin{aligned}
S_1 &:= \sum_{x=1}^{d} p_x |\beta_x| \,\mathsf{E}\,\mathrm{bayes}\big(N_x^p, |\beta_x|\big)\,\mathrm{I}\big\{N_x^p > (1-\varepsilon)m p_x\big\} \\
&= S_{11} + S_{12} \\
&\le \frac{(1 + A_1\sqrt{\varepsilon})c_\infty}{\sqrt{1-\varepsilon}}\sqrt{\frac{d}{m}}
\end{aligned} \tag{3.29}$$

for some universal real constant $A_1 > 0$.

On the other hand, by (3.25) and (3.26),

$$S_2 := \sum_{x=1}^{d} p_x |\beta_x| \,\mathsf{E}\, \text{bayes}\big(N_x^p, |\beta_x|\big) \,\mathrm{I}\big\{N_x^p \le (1-\varepsilon) m p_x\big\}$$
$$\le AS_{01} \le A a_\varepsilon \frac{d}{m} \le \varepsilon \sqrt{\frac{d}{m}}.$$

Combining this with (3.29), we conclude that

$$\sum_{x=1}^{d} p_x |\beta_x| \,\mathsf{E}\, \text{bayes}\big(N_x^p, |\beta_x|\big) = S_1 + S_2 \le \frac{(1 + A_2\sqrt{\varepsilon}) c_\infty}{\sqrt{1-\varepsilon}} \sqrt{\frac{d}{m}} \tag{3.30}$$

for some universal real constant $A_2 > 0$. Thus, letting here ε be arbitrarily small and recalling the definition of $B(m,d)$ in (2.13), we complete the proof of the asymptotic relation (3.17).

To complete the proof of Theorems 2.4 and 2.16, let us finally verify (3.21). If $k = 2j$ is even then, by (3.20),

$$b Q_0(k,b) = \binom{2j}{j} \frac{1}{4^j} b\big(1 - b^2\big)^j \le \frac{A_3}{\sqrt{k+1}} b e^{-b^2 k/2} \le \frac{A_4}{k+1}$$

for some universal real constants $A_3 > 0$ and $A_4 > 0$ and all $b \in [0,1]$; since $Q_0(k,b) = 0$ if k is odd, the above bound in fact holds for all $k = 0, 1, \dots$. So,

$$\sum_{x=1}^{d} p_x |\beta_x| \,\mathsf{E}\, Q_0\big(N_x^p, |\beta_x|\big) \le A_4 \sum_{x=1}^{d} p_x \,\mathsf{E}\, \frac{1}{N_x^p + 1} \le A_4 \bigg(a_\varepsilon + \frac{1}{1-\varepsilon}\bigg) \frac{d}{m};$$

the second inequality in the above display is obtained similar to the inequality in (3.28). Thus, (3.21) is verified, and the proof of Theorems 2.4 and 2.16 is complete. $\square$

PROOF OF THEOREM 2.1. Take any learning algorithm $L \in \mathcal{L}_{\mathsf{rand}}$. Take any $b \in (0,1]$ and then any $\varepsilon \in (0,b)$ and any $\beta \in \{-b,b\}^{[d]}$. Let D_β and Z_m^β be as in (2.22). Recall (2.6) and let $\hat{\Delta}^\beta := \Delta(L(Z_m^\beta, U), D^\beta)$, so that $\hat{\Delta}^\beta \le b$. So, $\mathrm{I}\{\hat{\Delta}^\beta > \varepsilon\} \ge \frac{1}{b-\varepsilon}(\hat{\Delta}^\beta - \varepsilon)$, whence $\mathsf{P}(\hat{\Delta}^\beta > \varepsilon) \ge \frac{1}{b-\varepsilon}(\mathsf{E}\,\hat{\Delta}^\beta - \varepsilon)$. Therefore, by (2.9), Theorem 2.6, Proposition 2.8, and Jensen's inequality,

$$\max_{\beta \in \{-b,b\}^{[d]}} \mathsf{P}\big(\hat{\Delta}^\beta > \varepsilon\big) \ge \frac{1}{2^d} \sum_{\beta \in \{-b,b\}^{[d]}} \mathsf{P}\big(\hat{\Delta}^\beta > \varepsilon\big)$$
$$\ge \frac{1}{b-\varepsilon} \bigg(\frac{1}{2^d} \sum_{\beta \in \{-b,b\}^{[d]}} \mathsf{E}\,\hat{\Delta}^\beta - \varepsilon\bigg)$$
$$\ge \frac{b\,\widetilde{\text{bayes}}(\nu, b) - \varepsilon}{b - \varepsilon}.$$

Take now any $\nu_* \in [3, \infty)$ and any real $\nu \geq \nu_*$. Then, by Lemma A.3 and (A.13), $\widetilde{\text{bayes}}(\nu, b) \geq \frac{1}{2}(1 - (\frac{i_{\nu_*}+1}{i_{\nu_*}})^{1/8}\psi_b(\nu))$. Further, take any $z \in (0, \sqrt{\nu_*}]$ and $w \in (0, z)$, and then take $b = z/\sqrt{\nu}$ and $\varepsilon = w/\sqrt{\nu}$, so that the conditions $b \in (0, 1]$ and $\varepsilon \in (0, b)$ assumed in the beginning of this proof do hold. Then

$$\max_{\beta \in \{-1,1\}^{[d]}} \mathsf{P}\left(\hat{\Delta}^\beta > \frac{w}{\sqrt{\nu}}\right) \geq \frac{1}{z - w}\left(\frac{z}{2}\left[1 - \left(\frac{i_{\nu_*}+1}{i_{\nu_*}}\right)^{1/8} \frac{\operatorname{erf}(z/\sqrt{2})}{\exp\{-z^2/(6\nu_*)\}}\right] - w\right)$$
$$=: P_{\mathsf{low}}(w, \nu_*, z).$$

It remains to note that $P_{\mathsf{low}}(\frac{1}{\sqrt{320}}, \frac{128}{10}, \frac{331}{1000}) > 0.238$, $P_{\mathsf{low}}(\frac{1}{\sqrt{320}}, 3, \frac{320}{1000}) > 0.227$, $P_{\mathsf{low}}(\frac{1}{\sqrt{413/10}}, \frac{128}{10}, \frac{681}{1000}) > 0.01563 > \frac{1}{64}$ and $P_{\mathsf{low}}(\frac{1}{\sqrt{496/10}}, 3, \frac{601}{1000}) > 0.0159 > \frac{1}{64}$. $\square$

APPENDIX A: IDENTITIES AND INEQUALITIES FOR BINOMIAL DISTRIBUTIONS: DETAILS CONCERNING THE FUNCTION bayes

Recall the definition of bayes in (2.14). Take any $b \in [0, 1]$ and $k \in \overline{1, \infty}$. By (2.15),

$$s_k(b) = \frac{1}{2^k}\sum_{i=0}^{j}\binom{k}{i}\left[(1+b)^{k-i}(1-b)^i - (1-b)^{k-i}(1+b)^i\right], \tag{A.1}$$

$$j := j_k := \lfloor k/2 \rfloor. \tag{A.2}$$

Using identities $(k-i)\binom{k}{i} = k\binom{k-1}{i}$ and $i\binom{k}{i} = k\binom{k-1}{i-1}$, we have

$$\begin{aligned}\frac{2^k}{k}s_k'(b) := &\sum_{i=0}^{j}\binom{k-1}{i}\left[(1+b)^{k-i-1}(1-b)^i + (1-b)^{k-i-1}(1+b)^i\right]\\ &- \sum_{i=1}^{j}\binom{k-1}{i-1}\left[(1+b)^{k-i}(1-b)^{i-1} + (1-b)^{k-i}(1+b)^{i-1}\right].\end{aligned}$$

Making in the second sum the substitution $i = r + 1$, then replacing there r back by i, and introducing

$$q := q_k := k - j - 1, \tag{A.3}$$

we have

$$\frac{2^k}{k}s_k'(b)\Big/\binom{k-1}{j} = (1+b)^q(1-b)^j + (1-b)^q(1+b)^j = 2\left(1-b^2\right)^q, \tag{A.4}$$

which is nonincreasing in $b \in [0, 1]$. So, the function s_k is concave.

Moreover, it follows that

$$s_k(b) = s_k'(0) S_q(b) \qquad \text{where } S_q(b) := \int_0^b (1-u^2)^q \, du. \tag{A.5}$$

For all $i \in \overline{0, \infty}$, in view of (A.2), (A.3) and (A.4),

$$q_{2i+2} = q_{2i+1} = i \quad \text{and} \quad s_{2i+2}'(0) = s_{2i+1}'(0) \tag{A.6}$$

and hence, by (A.5), one has the curious, and useful identity

$$s_{2i+2}(b) = s_{2i+1}(b). \tag{A.7}$$

We also have the following.

LEMMA A.1. *Take any $b \in (0, 1)$. Then the function*

$$\{0, 1, 3, 5, \dots\} \ni k \mapsto \text{bayes}(k, b) \tag{A.8}$$

is strictly convex.

PROOF. In view of (2.14), it is enough to show that the function $\{0, 1, 3, 5, \dots\} \ni k \mapsto s_k(b)$ is strictly concave. By (A.1) and (A.2),

$$s_0(b) = 0, \qquad s_1(b) = b, \qquad s_3(b) = \frac{1}{2}(3b - b^3) < 3s_1(b). \tag{A.9}$$

So, the restriction of the function in (A.8) to the set $\{0, 1, 3\}$ is strictly concave.

It remains to show that the restriction of this function to the set $\{1, 3, 5, \dots\}$ is strictly concave. Take any $i \in \overline{0, \infty}$. We have to show that

$$g(b) := \tilde{s}_i(b) + \tilde{s}_{i+2}(b) - 2\tilde{s}_{i+1}(b) < 0,$$

where

$$\tilde{s}_i := s_{2i+1}. \tag{A.10}$$

By (A.5), $\tilde{s}_\alpha'(b) = \tilde{s}_\alpha'(0)(1-b^2)^\alpha > 0$; here and in the rest of this proof, α stands for an arbitrary nonnegative integer. So, $g'(b)$ has the same sign as

$$\begin{aligned} &\frac{g'(b)}{\tilde{s}_{i+1}'(b)} 2(1-b^2)(i+2)(2i+3) \\ &\quad = -1 - 2(3+2i)w + (15 + 16i + 4i^2)w^2 =: g_1(w), \end{aligned} \tag{A.11}$$

where $w := b^2$. Since the function g_1 is convex, with $g_1(0) = -1 < 0$ and $g_1(1) = 4(2 + 3i + i^2) > 0$, it follows that $g_1(w)$ switches exactly once in sign, from $-$ to $+$, as w increases from 0 to 1. That is, $g(b)$ switches exactly once, from decreasing to increasing, as b increases from 0 to 1. Also, by (A.1) and (A.10), $\tilde{s}_\alpha(0) = 0$ and $\tilde{s}_\alpha(1) = 1$, where $g(0) = 0 = g(1)$. Thus, indeed $g(b) < 0$ for all $b \in (0, 1)$. $\square$

LEMMA A.2. *For $i \in \overline{0,\infty}$, let*

$$C_i := \sqrt{\frac{\pi}{2}}\frac{s'_{2i+1}(0)}{\sqrt{2i+1}} = \frac{\sqrt{\pi(i+1/2)}}{2^{2i}}\binom{2i}{i}, \tag{A.12}$$

the latter equality following by (A.4) *and* (A.2).

Then C_i decreases from $\sqrt{\frac{\pi}{2}}$ to 1 as i increases from 0 to ∞, and for all $i \in \overline{1,\infty}$

$$C_i < \left(\frac{i+1}{i}\right)^{1/8}. \tag{A.13}$$

PROOF. In this proof, it is assumed that $i \in \overline{0,\infty}$. Let

$$r_i := \frac{C_i}{C_{i+1}} = \frac{2i+2}{\sqrt{(2i+2)^2-1}} > 1.$$

So, C_i indeed decreases in i. It is easy to check that $C_0 = \sqrt{\frac{\pi}{2}}$ and $C_i \to C_\infty := 1$ as $i \to \infty$.

It remains to verify inequality (A.13). Accordingly, assume through the end of this proof that $i \in \overline{1,\infty}$. Then

$$-1 + r_i^{-8}\Big/\left(1 - \frac{1}{(i+1)^2}\right) = \frac{96i^4 + 384i^3 + 560i^2 + 352i + 81}{256i(i+1)^6(i+2)} > 0,$$

which shows that

$$r_i < \left(1 - \frac{1}{(i+1)^2}\right)^{-1/8},$$

where

$$\begin{aligned} C_i &= C_\infty \prod_{\alpha=i}^{\infty} r_\alpha = \prod_{\alpha=i}^{\infty} r_\alpha < \prod_{\alpha=i}^{\infty}\left(1 - \frac{1}{(\alpha+1)^2}\right)^{-1/8} \\ &= \prod_{\alpha=i}^{\infty}\left(\frac{\alpha}{\alpha+1}\Big/\frac{\alpha+1}{\alpha+2}\right)^{-1/8} = \left(\frac{i+1}{i}\right)^{1/8}, \end{aligned}$$

which completes the proof of Lemma A.2. □

LEMMA A.3. *Take any real $\kappa \ge 1$ and any $b \in [0,1]$. Then*

$$\widetilde{\text{bayes}}(\kappa, b) \ge \frac{1}{2}\big(1 - C_{i_\kappa}\psi_b(\kappa)\big),$$

where $i_\kappa := \lfloor\frac{\kappa-1}{2}\rfloor$, C_i as in (A.12), *and*

$$\psi_b(\kappa) := e^{b^2/6}\operatorname{erf}(b\sqrt{\kappa/2}).$$

PROOF. For brevity, let $i := i_\kappa = \lfloor \frac{\kappa-1}{2} \rfloor$ and $k := 2i+1$. Then $i \in \overline{0,\infty}$, $k = 2i+1 \le \kappa < 2i+3 = k+2$, and so, by (2.25),

$$\widetilde{\text{bayes}}(\kappa, b) = \frac{k+2-\kappa}{2}\,\text{bayes}(k,b) + \frac{\kappa-k}{2}\,\text{bayes}(k+2,b). \tag{A.14}$$

In view of (A.2) and (A.3), $j_k = i = q_k = q$. It is well known and easily proved (cf., e.g., [13]) that all mixtures of log-convex functions are log-convex. Therefore, and because the function $q \mapsto e^{-qu^2}$ is log-convex, $\tilde{S}_q(b) := \int_0^b e^{-qu^2}\,du$ is log-convex in real q; that is, $f(q) := \ln \tilde{S}_q(b)$ is convex in q. [For readers' convenience, a quick and direct way to verify the convexity of f is to see that for any real q_1, q_2 and any $t \in (0,1)$ the inequality $f((1-t)q_1 + tq_2) \le (1-t)f(q_1) + tf(q_2)$ can be rewritten as the instance of Hölder's inequality $\int_0^b F_1(u)F_2(u)\,du \le (\int_0^b F_1(u)^p\,du)^{1/p}(\int_0^b F_2(u)^{\tilde{p}}\,du)^{1/\tilde{p}}$ with $F_1(u) := e^{-(1-t)q_1u^2}$, $F_2(u) := e^{-tq_2u^2}$, $p := \frac{1}{1-t}$, and $\tilde{p} := \frac{1}{t}$.] So, $\ln \frac{\tilde{S}_{q+1/2}(b)\tilde{S}_0(b)}{\tilde{S}_{1/2}(b)\tilde{S}_q(b)} = f(q+1/2) - f(q) - f(1/2) + f(0) = \int_0^{1/2}[f'(q+u) - f'(u)]\,du \ge 0$ for $q \ge 0$, since $f'(u)$ is increasing in u (instead of the last equality here, one can also use inequality [8], (3.17.5), with $\phi = f$, $h = x = q/2 + 1/4$ and $h' = q/2 - 1/4$ there). Recalling now (A.5) and using the elementary inequality $1 - t \le e^{-t}$ for real t, we have

$$\begin{aligned}
S_q(b) &\le \tilde{S}_q(b) \le \tilde{S}_{q+1/2}(b)\frac{\tilde{S}_0(b)}{\tilde{S}_{1/2}(b)} \\
&= \frac{\sqrt{\pi}}{2}\frac{\operatorname{erf}(b\sqrt{q+1/2})}{\sqrt{q+1/2}}\frac{b}{\int_0^b e^{-u^2/2}\,du} \\
&\le \frac{\sqrt{\pi}}{2}\frac{\operatorname{erf}(b\sqrt{q+1/2})}{\sqrt{q+1/2}}e^{b^2/6} \\
&= \sqrt{\frac{\pi}{2}}\frac{1}{\sqrt{k}}e^{b^2/6}\operatorname{erf}(b\sqrt{k/2}) = \sqrt{\frac{\pi}{2}}\frac{1}{\sqrt{k}}\psi_b(k); \qquad \text{(A.15)}
\end{aligned}$$

here we used the inequality $g(b) := \int_0^b e^{-u^2/2}\,du - be^{-b^2/6} > 0$ for $b > 0$, which follows because $g(0) = 0$ and $g'(b) = e^{-b^2/6}(e^{-b^2/3} - (1 - b^2/3)) > 0$ for $b > 0$. So, in view of (2.14), (A.5) and (A.12), $\text{bayes}(k,b) \ge \frac{1}{2}(1 - C_i\psi_b(k))$. Replacing here k by $k+2$, one has

$$\text{bayes}(k+2,b) \ge \frac{1}{2}\big(1 - C_{i+1}\psi_b(k+2)\big) \ge \frac{1}{2}\big(1 - C_i\psi_b(k+2)\big);$$

the last inequality here follows because, by Lemma A.2, C_i decreases in i. Now (A.14) yields

$$\widetilde{\text{bayes}}(\kappa,b) \ge \frac{1}{2} - \frac{1}{2}C_i\left[\frac{k+2-\kappa}{2}\psi_b(k) + \frac{\kappa-k}{2}\psi_b(k+2)\right] \ge \frac{1}{2} - \frac{1}{2}C_i\psi_b(\kappa),$$

the latter inequality following by the concavity of $\psi_b(u)$ in $u \ge 0$. This completes the proof of Lemma A.3. $\square$

APPENDIX B: SIMPLIFIED FORM OF $c^{\mathsf{LB}}_{m,d}$

Take any finite set $\mathcal{X}$. Take then any hypothesis class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = d$. Let now $\tilde{\mathcal{X}}$ be any subset of $\mathcal{X}$ of cardinality d such that $\tilde{\mathcal{X}}$ is shattered by $\mathcal{H}$. Clearly, for any learning algorithm $L \in \mathcal{L}_{\mathcal{X};m,\mathcal{H}}$ we have

$$\sup_{D \in \mathcal{D}_{\mathcal{X}}} \mathfrak{R}_m(L, D) \ge \sup_{D \in \mathcal{D}_{\mathcal{X};\tilde{\mathcal{X}}}} \mathfrak{R}_m(L, D), \tag{B.1}$$

where $D_{\mathcal{X};\tilde{\mathcal{X}}}$ is the set of all distributions $D \in \mathcal{D}_{\mathcal{X}}$ with support contained in the set $\tilde{\mathcal{X}} \times \mathcal{Y}$ (and, recall, $\mathcal{D}_{\mathcal{X}}$ is the set of all distributions on $\mathcal{X} \times \mathcal{Y}$). Next, any distribution $D \in \mathcal{D}_{\mathcal{X};\tilde{\mathcal{X}}}$ may be identified with the corresponding distribution, say $\tilde{D}$, in $\mathcal{D}_{\tilde{\mathcal{X}}}$. Also, for any $D \in \tilde{\mathcal{D}}_{\mathcal{X};\tilde{\mathcal{X}}}$, the value of $\mathfrak{R}_m(L, D)$ depends on L only through its restriction, say $\tilde{L}$, to $(\tilde{\mathcal{X}} \times \mathcal{Y})^k$ so that $\mathfrak{R}_m(L, D) = \mathfrak{R}_m(\tilde{L}, \tilde{D})$. Moreover, we may identify the finite set $\mathcal{X}$ with the set $[N] = \{1, \dots, N\}$ for some natural $N \ge d$, and then without loss of generality $\tilde{\mathcal{X}} = [d] = \{1, \dots, d\}$. Thus, for any finite set $\mathcal{X}$ and any hypothesis class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = d$, it follows from (B.1) that

$$\inf_{L \in \mathcal{L}_{\mathcal{X};m,\mathcal{H}}} \sup_{D \in \mathcal{D}_{\mathcal{X}}} \mathfrak{R}_m(L, D) \ge \inf_{\tilde{L} \in \mathcal{L}_{[d];m,\mathcal{Y}^{[d]}}} \sup_{\tilde{D} \in \mathcal{D}_{[d]}} \mathfrak{R}_m(\tilde{L}, \tilde{D}). \tag{B.2}$$

The right-hand side (RHS) of (B.2) does not depend on $\mathcal{X}$ or $\mathcal{H}$, as long as $\mathrm{VC}(\mathcal{H}) = d$. So, by (1.10), $c^{\mathsf{LB}}_{m,d}/\sqrt{m/d} \ge$ RHS of (B.2). The reverse of the latter inequality follows trivially from (1.10), since the set $[d]$ is finite and $\mathrm{VC}(\mathcal{Y}^{[d]}) = d$. We conclude that $c^{\mathsf{LB}}_{m,d}/\sqrt{m/d} =$ RHS of (B.2). Thus, (2.2) follows.

Acknowledgments. We are pleased to thank Peter Grünwald for bringing the result of [2] to our attention, and the referees for carefully reading the paper and providing useful suggestions on the presentation.

REFERENCES

[1] ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge Univ. Press, Cambridge. MR1741038

[2] AUDIBERT, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *Ann. Statist.* **37** 1591–1646. MR2533466

[3] BEREND, D. and KONTOROVICH, A. (2015). A finite sample analysis of the naive Bayes classifier. *J. Mach. Learn. Res.* **16** 1519–1545. MR3417789

[4] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities. A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193

[5] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. Springer, New York. MR1383093

[6] DEVROYE, L. and LUGOSI, G. (1995). Lower bounds in pattern recognition and learning. *Pattern Recognit*. **28** 1011–1018.

[7] FERGUSON, T. S. (1967). *Mathematical Statistics*: *A Decision Theoretic Approach*. *Probability and Mathematical Statistics* **1**. Academic Press, New York. MR0215390

[8] HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1988). *Inequalities*. Cambridge Univ. Press, Cambridge. MR0944909

[9] HAUSSLER, D. (1992). Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Inform*. *and Comput*. **100** 78–150. MR1175977

[10] HAUSSLER, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. *J. Combin*. *Theory Ser. A* **69** 217–232. MR1313896

[11] KEARNS, M. J. and SCHAPIRE, R. E. (1994). Efficient distribution-free learning of probabilistic concepts. *J. Comput*. *System Sci*. **48** 464–497. MR1279411

[12] KEARNS, M. J., SCHAPIRE, R. E. and SELLIE, L. (1994). Toward efficient agnostic learning. *Mach*. *Learn*. **17** 115–141.

[13] KINGMAN, J. F. C. (1961). A convexity property of positive matrices. *Quart*. *J. Math*. *Oxford Ser.* (2) **12** 283–284. MR0138632

[14] KONTOROVICH, A., SABATO, S. and URNER, R. (2016). Active nearest-neighbor learning in metric spaces. In *NIPS 2016*. Available at arXiv:1605.06792.

[15] LONG, P. M. (1999). The complexity of learning according to two models of a drifting environment. *Mach*. *Learn*. **37** 337–354.

[16] PINELIS, I. (2016). Optimal binomial, Poisson, and normal left-tail domination for sums of nonnegative random variables. *Electron*. *J. Probab.* **21** Paper No. 20. MR3485362

[17] PINELIS, I. F. (1991). Criteria for complete determinateness for concave-convex games. *Mat*. *Zametki* **49** 73–76, 159. MR1110310

[18] PINELIS, I. S. and UTEV, S. A. (1989). Sharp exponential estimates for sums of independent random variables. *Teor. Veroyatn*. *Primen*. **34** 384–390. MR1005745

[19] SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014). *Understanding Machine Learning*: *From Theory to Algorithms*. Cambridge Univ. Press, Cambridge.

[20] SIMON, H. U. (1996). General bounds on the number of examples needed for learning probabilistic concepts. *J. Comput*. *System Sci*. **52** 239–254. MR1393992

[21] SION, M. (1958). On general minimax theorems. *Pacific J. Math*. **8** 171–176. MR0097026

[22] TALAGRAND, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann*. *Probab.* **22** 28–76. MR1258865

[23] V. NEUMANN, J. (1928). Zur Theorie der Gesellschaftsspiele. *Math*. *Ann*. **100** 295–320. MR1512486

[24] VALIANT, L. G. (1984). A theory of the learnable. *Commun*. *ACM* **27** 1134–1142.

[25] VAPNIK, V. N. and ČERVONENKIS, A. JA. (1971). The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Veroyatn*. *Primen*. **16** 264–279. MR0288823

DEPARTMENT OF COMPUTER SCIENCE
BEN-GURION UNIVERSITY
BEER SHEVA 84105
ISRAEL
E-MAIL: karyeh@cs.bgu.ac.il

DEPARTMENT OF MATHEMATICAL SCIENCES
MICHIGAN TECHNOLOGICAL UNIVERSITY
HOUGHTON, MICHIGAN 49931-1295
USA
E-MAIL: ipinelis@mtu.edu