Journal of Machine Learning Research 16 (2015) 1519-1545 Submitted 8/14; Revised 11/14; Published 8/15

A Finite Sample Analysis of the Naive Bayes Classifier*

Daniel Berend BEREND@CS.BGU.AC.IL
Department of Computer Science and Department of Mathematics
Ben-Gurion University
Beer Sheva, Israel

Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Department of Computer Science
Ben-Gurion University
Beer Sheva, Israel

Editor: Gabor Lugosi

Abstract

We revisit, from a statistical learning perspective, the classical decision-theoretic problem of weighted expert voting. In particular, we examine the consistency (both asymptotic and finitary) of the optimal Naive Bayes weighted majority and related rules. In the case of known expert competence levels, we give sharp error estimates for the optimal rule. We derive optimality results for our estimates and also establish some structural characterizations. When the competence levels are unknown, they must be empirically estimated. We provide frequentist and Bayesian analyses for this situation. Some of our proof techniques are non-standard and may be of independent interest. Several challenging open problems are posed, and experimental results are provided to illustrate the theory.

Keywords: experts, hypothesis testing, Chernoff-Stein lemma, Neyman-Pearson lemma, naive Bayes, measure concentration

1. Introduction

Imagine independently consulting a small set of medical experts for the purpose of reaching a binary decision (e.g., whether to perform some operation). Each doctor has some "reputation", which can be modeled as his probability of giving the right advice. The problem of weighting the input of several experts arises in many situations and is of considerable theoretical and practical importance. The rigorous study of majority vote has its roots in the work of Condorcet (1785). By the 70s, the field of decision theory was actively exploring various voting rules (see Nitzan and Paroush (1982) and the references therein). A typical setting is as follows. An agent is tasked with predicting some random variable $Y \in \{\pm 1\}$ based on input $X_i \in \{\pm 1\}$ from each of n experts. Each expert X_i has a *competence* level $p_i \in (0,1)$, which is his probability of making a correct prediction: $\mathbb{P}(X_i = Y) = p_i$. Two simplifying assumptions are commonly made:

*. An extended abstract of this paper appeared in NIPS 2014 under the title "Consistency of weighted majority votes," which was also the former title of this paper. A. Kontorovich was partially supported by the Israel Science Foundation (grant No. 1141/12) and a Yahoo Faculty award.

(i) *Independence*: The random variables $\{X_i : i \in [n]\}$ are mutually independent conditioned on the truth Y.

(ii) *Unbiased truth*: $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1) = 1/2$.

We will discuss these assumptions below in greater detail; for now, let us just take them as given. (Since the bias of Y can be easily estimated from data, and the generalization to the asymmetric case is straightforward, only the independence assumption is truly restrictive.) A *decision rule* is a mapping $f : \{\pm 1\}^n \to \{\pm 1\}$ from the n expert inputs to the agent's final decision. Our quantity of interest throughout the paper will be the agent's probability of error,

$$\mathbb{P}(f(\mathbf{X}) \neq Y). \tag{1}$$

A decision rule f is *optimal* if it minimizes the quantity in (1) over all possible decision rules. It follows from the work of Neyman and Pearson (1933) that, when Assumptions (i)–(ii) hold and the true competences p_i are known, the optimal decision rule is obtained by an appropriately weighted majority vote:

$$f^{\mathrm{OPT}}(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i x_i\right), \tag{2}$$

where the weights w_i are given by

$$w_i = \log \frac{p_i}{1 - p_i}, \qquad i \in [n]. \tag{3}$$

Thus, w_i is the log-odds of expert i being correct, and the voting rule in (2) is also known as *naive Bayes* (Hastie et al., 2009).

Main results. Formula (2) raises immediate questions, which apparently have not previously been addressed. The first one has to do with the *consistency* of the naive Bayes decision rule: under what conditions does the probability of error decay to zero and at what rate? In Section 3, we show that the probability of error is controlled by the *committee potential* Φ, defined by

$$\Phi = \sum_{i=1}^{n} (p_i - \tfrac{1}{2}) w_i = \sum_{i=1}^{n} (p_i - \tfrac{1}{2}) \log \frac{p_i}{1 - p_i}. \tag{4}$$

More precisely, we prove in Theorem 1 that

$$-\log \mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) \neq Y) \quad \asymp \quad \Phi,$$

where $\asymp$ denotes equivalence up to universal multiplicative constants. As we show in Section 3.3, both the upper estimate of $O(e^{-\Phi/2})$ and the lower one of $\Omega(e^{-2\Phi})$ are tight in various regimes of Φ. The structural characterization in terms of "antipodes" (Lemma 2) and the additional bounds provided in Section 3.4 may also be of interest.

Another issue not addressed by the Neyman-Pearson lemma is how to handle the case where the competences p_i are not known exactly but rather estimated empirically by $\hat{p}_i$. We

present two solutions to this problem: a frequentist and a Bayesian one. As we show in Section 4, the frequentist approach does not admit an optimal empirical decision rule. Instead, we analyze empirical decision rules in various settings: high-confidence (i.e., $|\hat{p}_i - p_i| \ll 1$) vs. low-confidence, adaptive vs. nonadaptive. The low-confidence regime requires no additional assumptions, but gives weaker guarantees (Theorem 7). In the high-confidence regime, the adaptive approach produces error estimates in terms of the empirical $\hat{p}_i$s (Theorem 13), while the nonadaptive approach yields a bound in terms of the unknown p_is, which still leads to useful asymptotics (Theorem 11). The Bayesian solution sidesteps the various cases above, as it admits a simple, provably optimal empirical decision rule (Section 5). Unfortunately, we are unable to compute (or even nontrivially estimate) the probability of error induced by this rule; this is posed as a challenging open problem.

Notation. We use standard set-theoretic notation, and in particular $[n] = \{1, \ldots, n\}$.

2. Related Work

The Naive Bayes weighted majority voting rule was stated by Nitzan and Paroush (1982) in the context of decision theory, but its roots trace much earlier to the problem of hypothesis testing (Neyman and Pearson, 1933). Machine learning theory typically clusters *weighted majority* (Littlestone and Warmuth, 1989, 1994) within the framework of online algorithms; see Cesa-Bianchi and Lugosi (2006) for a modern treatment. Since the online setting is considerably more adversarial than ours, we obtain very different weighted majority rules and consistency guarantees. The weights w_i in (2) bear a striking similarity to the AdaBoost update rule (Freund and Schapire, 1997; Schapire and Freund, 2012). However, the latter assumes weak learners with access to labeled examples, while in our setting the experts are "static". Still, we do not rule out a possible deeper connection between the Naive Bayes decision rule and boosting.

In what began as the influential Dawid-Skene model (Dawid and Skene, 1979) and is now known as *crowdsourcing*, one attempts to extract accurate predictions by pooling a large number of experts, typically without the benefit of being able to test any given expert's competence level. Still, under mild assumptions it is possible to efficiently recover the expert competences to a high accuracy and to aggregate them effectively (Parisi et al., 2014+). Error bounds for the oracle MAP rule were obtained in this model by Li et al. (2013) and minimax rates were given in Gao and Zhou (2014).

In a recent line of work, Lacasse et al. (2006); Laviolette and Marchand (2007); Roy et al. (2011) have developed a PAC-Bayesian theory for the majority vote of simple classifiers. This approach facilitates data-dependent bounds and is even flexible enough to capture some simple dependencies among the classifiers — though, again, the latter are *learners* as opposed to our *experts*. Even more recently, experts with adversarial noise have been considered (Mansour et al., 2013), and efficient algorithms for computing optimal expert weights (without error analysis) were given (Eban et al., 2014). More directly related to the present work are the papers of Berend and Paroush (1998), which characterizes the conditions for the consistency of the simple majority rule, and Boland et al. (1989); Berend and Sapir (2007); Helmbold and Long (2012) which analyze various models of dependence among the experts.

3. Known Competences

In this section we assume that the expert competences p_i are known and analyze the consistency of the naive Bayes decision rule (2). Our main result here is that the probability of error $\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y)$ is small if and only if the committee potential Φ is large.

Theorem 1 *Suppose that the experts $\mathbf{X} = (X_1, \ldots, X_n)$ satisfy Assumptions (i)-(ii) and $f^{\text{OPT}} : \{\pm 1\}^n \to \{\pm 1\}$ is the naive Bayes decision rule in (2). Then*

(i) $\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) \leq \exp\left(-\frac{1}{2}\Phi\right)$.

(ii) $\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) \geq \dfrac{3}{4[1+\exp(2\Phi + 4\sqrt{\Phi})]}$.

The next two sections are devoted to proving Theorem 1. These are followed by an optimality result and some additional upper and lower bounds.

3.1 Proof of Theorem 1(i)

Define the $\{0,1\}$-indicator variables

$$\xi_i = \mathbb{1}_{\{X_i = Y\}}, \tag{5}$$

corresponding to the event that the i^{th} expert is correct. A mistake $f^{\text{OPT}}(\mathbf{X}) \neq Y$ occurs precisely when[1] the sum of the correct experts' weights fails to exceed half the total mass:

$$\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) = \mathbb{P}\left(\sum_{i=1}^n w_i \xi_i \leq \frac{1}{2}\sum_{i=1}^n w_i\right). \tag{6}$$

Since $\mathbb{E}\xi_i = p_i$, we may rewrite the probability in (6) as

$$\mathbb{P}\left(\sum_i w_i \xi_i \leq \mathbb{E}\left[\sum_i w_i \xi_i\right] - \sum_i (p_i - \tfrac{1}{2}) w_i\right). \tag{7}$$

A standard tool for estimating such sum deviation probabilities is Hoeffding's inequality (Hoeffding, 1963). Applied to (7), it yields the bound

$$\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) \leq \exp\left(-\frac{2\left[\sum_i (p_i - \frac{1}{2}) w_i\right]^2}{\sum_i w_i^2}\right), \tag{8}$$

which is far too crude for our purposes. Indeed, consider a finite committee of highly competent experts with p_i's arbitrarily close to 1 and X_1 the most competent of all. Raising X_1's competence sufficiently far above his peers will cause both the numerator and the denominator in the exponent to be dominated by w_1^2, making the right-hand-side of (8) bounded away from zero. In the limiting case of this regime, the probability of error approaches zero while the right-hand side of (8) approaches $e^{-1/2} \approx 0.6$. The inability of Hoeffding's inequality to guarantee consistency even in such a felicitous setting is an instance

1. Without loss of generality, ties are considered to be errors.

of its generally poor applicability to highly heterogeneous sums, a phenomenon explored in some depth in McAllester and Ortiz (2003). Bernstein's and Bennett's inequalities suffer from a similar weakness (see ibid.). Fortunately, an inequality of Kearns and Saul (1998) is sufficiently sharp[2] to yield the desired estimate: For all $p \in [0,1]$ and all $t \in \mathbb{R}$,

$$(1-p)e^{-tp} + pe^{t(1-p)} \leq \exp\left(\frac{1-2p}{4\log((1-p)/p)}t^2\right). \tag{9}$$

Put $\theta_i = \xi_i - p_i$, substitute into (6), and apply Markov's inequality:

$$\begin{aligned} \mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) \neq Y) &= \mathbb{P}\left(-\sum_i w_i\theta_i \geq \Phi\right) \\ &\leq e^{-t\Phi}\mathbb{E}\exp\left(-t\sum_i w_i\theta_i\right). \end{aligned} \tag{10}$$

Now

$$\begin{aligned} \mathbb{E}e^{-tw_i\theta_i} &= p_i e^{-(1-p_i)w_i t} + (1-p_i)e^{p_i w_i t} \\ &\leq \exp\left(\frac{-1+2p_i}{4\log(p_i/(1-p_i))}w_i^2 t^2\right) \\ &= \exp\left[\tfrac{1}{2}(p_i - \tfrac{1}{2})w_i t^2\right], \end{aligned} \tag{11}$$

where the inequality follows from (9). By independence,

$$\begin{aligned} \mathbb{E}\exp\left(-t\sum_i w_i\theta_i\right) &= \prod_i \mathbb{E}e^{-tw_i\theta_i} \\ &\leq \exp\left(\tfrac{1}{2}\sum_i (p_i - \tfrac{1}{2})w_i t^2\right) \\ &= \exp\left(\tfrac{1}{2}\Phi t^2\right) \end{aligned}$$

and hence

$$\mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) \neq Y) \leq \exp\left(\tfrac{1}{2}\Phi t^2 - \Phi t\right).$$

Choosing $t = 1$, we obtain the bound in Theorem 1(i).

3.2 Proof of Theorem 1(ii)

Define the $\{\pm 1\}$-indicator variables

$$\eta_i = 2 \cdot \mathbb{1}_{\{X_i = Y\}} - 1, \tag{12}$$

corresponding to the event that the i^{th} expert is correct, and put $q_i = 1 - p_i$. The shorthand $\mathbf{w} \cdot \boldsymbol{\eta} = \sum_{i=1}^n w_i\eta_i$ will be convenient. We will need some simple lemmata:

2. The Kearns-Saul inequality (9) may be seen as a distribution-dependent refinement of Hoeffding's for a two-valued distribution (which bounds the left-hand-side of (9) by $e^{t^2/8}$), and is not nearly as straightforward to prove; see Appendix A.

Lemma 2

$$\begin{aligned}\mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) = Y) &= \tfrac{1}{2} \sum_{\boldsymbol{\eta} \in \{\pm 1\}^n} \max\left\{P(\boldsymbol{\eta}), P(-\boldsymbol{\eta})\right\} \\ &= \sum_{\boldsymbol{\eta} \in \{+1\} \times \{\pm 1\}^{n-1}} \max\left\{P(\boldsymbol{\eta}), P(-\boldsymbol{\eta})\right\}\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) \neq Y) &= \tfrac{1}{2} \sum_{\boldsymbol{\eta} \in \{\pm 1\}^n} \min\left\{P(\boldsymbol{\eta}), P(-\boldsymbol{\eta})\right\} \\ &= \sum_{\boldsymbol{\eta} \in \{+1\} \times \{\pm 1\}^{n-1}} \min\left\{P(\boldsymbol{\eta}), P(-\boldsymbol{\eta})\right\},\end{aligned}$$

where

$$P(\boldsymbol{\eta}) = \prod_{i:\eta_i=1} p_i \prod_{i:\eta_i=-1} q_i. \tag{13}$$

Proof By (5), (6) and (12), that a mistake occurs precisely when

$$\sum_{i=1}^n w_i \frac{\eta_i + 1}{2} \leq \frac{1}{2} \sum_{i=1}^n w_i,$$

which is equivalent to

$$\mathbf{w} \cdot \boldsymbol{\eta} \leq 0. \tag{14}$$

Exponentiating both sides,

$$\begin{aligned}\exp\left(\mathbf{w} \cdot \boldsymbol{\eta}\right) &= \prod_{i=1}^n e^{w_i \eta_i} \\ &= \prod_{i:\eta_i=1} \frac{p_i}{q_i} \cdot \prod_{i:\eta_i=-1} \frac{q_i}{p_i} \\ &= \frac{P(\boldsymbol{\eta})}{P(-\boldsymbol{\eta})} \;\leq\; 1. \end{aligned} \tag{15}$$

We conclude from (15) that among two "antipodal" atoms $\pm\boldsymbol{\eta} \in \{\pm 1\}^n$, the one with the greater mass contributes to the probability of being correct and the one with the smaller mass contributes to the probability of error, which proves the claim. ■

Lemma 3 *Suppose that* $\mathbf{s}, \mathbf{s}' \in (0, \infty)^m$ *satisfy*

$$\sum_{i=1}^m (s_i + s_i') \geq a$$

and

$$\frac{1}{R} \leq \frac{s_i}{s_i'} \leq R, \qquad i \in [m]$$

for some $1 \leq R < \infty$. *Then*

$$\sum_{i=1}^{m} \min\left\{s_i, s_i'\right\} \geq \frac{a}{1+R}.$$

Proof Immediate from

$$s_i + s_i' \leq \min\left\{s_i, s_i'\right\}(1+R).$$

■

Lemma 4 *Define the function* $F : (0,1) \to \mathbb{R}$ *by*

$$F(x) = \frac{x(1-x)\log(x/(1-x))}{2x-1}.$$

Then $\sup_{0<x<1} F(x) = \frac{1}{2}$.

Proof Since F is symmetric about $x = \frac{1}{2}$, it suffices to prove the claim for $\frac{1}{2} \leq x < 1$. We will show that F is concave by examining its second derivative:

$$F''(x) = -\frac{2x - 1 - 2x(1-x)\log(x/(1-x))}{x(1-x)(2x-1)^3}.$$

The denominator is obviously nonnegative on $[\frac{1}{2}, 1]$, while the numerator has the Taylor expansion

$$\sum_{n=1}^{\infty} \frac{2^{2(n+1)}(x-\frac{1}{2})^{2n+1}}{4n^2-1} \geq 0, \qquad \frac{1}{2} \leq x < 1$$

(verified through tedious but straightforward calculus). Since F is concave and symmetric about $\frac{1}{2}$, its maximum occurs at $F(\frac{1}{2}) = \frac{1}{2}$. ■

Continuing with the main proof, observe that

$$\mathbb{E}\left[\mathbf{w} \cdot \boldsymbol{\eta}\right] = \sum_{i=1}^{n} (p_i - q_i) w_i = 2\Phi \tag{16}$$

and

$$\mathrm{Var}\left[\mathbf{w} \cdot \boldsymbol{\eta}\right] = 4 \sum_{i=1}^{n} p_i q_i w_i^2.$$

By Lemma 4,

$$p_i q_i w_i^2 \leq \tfrac{1}{2}(p_i - q_i) w_i,$$

and hence

$$\mathrm{Var}\left[\mathbf{w} \cdot \boldsymbol{\eta}\right] \;\; \leq \;\; 4\Phi. \tag{17}$$

Define the segments $I, J \subset \mathbb{R}$ by

$$I = \left[2\Phi - 4\sqrt{\Phi}, 2\Phi + 4\sqrt{\Phi}\right] \subset \left[-2\Phi - 4\sqrt{\Phi}, 2\Phi + 4\sqrt{\Phi}\right] = J. \tag{18}$$

Chebyshev's inequality together with (16, 17, 18) implies that

$$\mathbb{P}\left(\mathbf{w} \cdot \boldsymbol{\eta} \in J\right) \;\geq\; \mathbb{P}\left(\mathbf{w} \cdot \boldsymbol{\eta} \in I\right) \;\geq\; \frac{3}{4}. \tag{19}$$

Consider an atom $\boldsymbol{\eta} \in \{\pm 1\}^n$ for which $\mathbf{w} \cdot \boldsymbol{\eta} \in J$. It follows from (15) and (18) that

$$\frac{P(\boldsymbol{\eta})}{P(-\boldsymbol{\eta})} = \exp\left(\mathbf{w} \cdot \boldsymbol{\eta}\right) \leq \exp(2\Phi + 4\sqrt{\Phi}). \tag{20}$$

Finally, we have

$$\begin{aligned}
\mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) \neq Y) \;\; &\overset{(a)}{=} \sum_{\boldsymbol{\eta} \in \{+1\} \times \{\pm 1\}^{n-1}} \min\left\{P(\boldsymbol{\eta}), P(-\boldsymbol{\eta})\right\} \\
&\geq \sum_{\boldsymbol{\eta} \in \{+1\} \times \{\pm 1\}^{n-1} : \mathbf{w} \cdot \boldsymbol{\eta} \in J} \min\left\{P(\boldsymbol{\eta}), P(-\boldsymbol{\eta})\right\} \\
&\overset{(b)}{\geq} \frac{1}{1 + \exp(2\Phi + 4\sqrt{\Phi})} \sum_{\boldsymbol{\eta} \in \{+1\} \times \{\pm 1\}^{n-1} : \mathbf{w} \cdot \boldsymbol{\eta} \in J} \left(P(\boldsymbol{\eta}) + P(-\boldsymbol{\eta})\right) \\
&\overset{(c)}{=} \frac{1}{1 + \exp(2\Phi + 4\sqrt{\Phi})} \sum_{\boldsymbol{\eta} \in \{\pm 1\}^{n} : \mathbf{w} \cdot \boldsymbol{\eta} \in J} P(\boldsymbol{\eta}) \\
&\overset{(d)}{\geq} \frac{3/4}{1 + \exp(2\Phi + 4\sqrt{\Phi})},
\end{aligned}$$

where: (a) follows from Lemma 2, (b) from Lemma 3 and (20), (c) from the fact that $\mathbf{w} \cdot \boldsymbol{\eta} \in J \iff -\mathbf{w} \cdot \boldsymbol{\eta} \in J$, and (d) from (19). This completes the proof.

Remark 5 *The constant $\frac{3}{4}$ can be made arbitrarily close to 1 at the expense of an increased coefficient in front of the $\sqrt{\Phi}$ term. More precisely, the $4\sqrt{\Phi}$ term in (18) corresponds to taking two standard deviations about the mean. Taking instead k standard deviations would cause $4\sqrt{\Phi}$ to be replaced by $2k\sqrt{\Phi}$ and the $\frac{3}{4}$ constant to be replaced by $1 - 1/k^2$. This leads to (mild) improvements for large Φ.*

3.3 Asymptotic tightness

Although there is a 4^{th} power gap between the upper bound $U = \exp\left(-\frac{1}{2}\Phi\right)$ and lower bound $L \asymp \exp(-2\Phi)$ in Theorem 1, we will show that each estimate is tight in a certain regime of Φ.

Upper bound. To establish the tightness of the upper bound $U = e^{-\Phi/2}$, consider n identical experts with competences $p_1 = \ldots = p_n = p > \frac{1}{2}$. Then

$$\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) = \mathbb{P}(B < \tfrac{1}{2}n) = \mathbb{P}(B < n(p-\varepsilon)), \tag{21}$$

where $B \sim \text{Bin}(n,p)$ and $\varepsilon = p - \frac{1}{2}$. By Sanov's theorem (den Hollander, 2000),

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(B < n(p-\varepsilon)) = H(p-\varepsilon||p) = H(\tfrac{1}{2}||p), \tag{22}$$

where

$$H(x||y) = x\ln\frac{x}{y} + (1-x)\ln\frac{1-x}{1-y}, \qquad 0 < x, y < 1.$$

Hence,

$$\begin{aligned} \frac{1}{n}\log\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) &\overset{(a)}{=} \frac{1}{n}\log\mathbb{P}(B < \tfrac{1}{2}n) \\ &\xrightarrow[n\to\infty]{(b)} -H(\tfrac{1}{2}||p) \\ &= \tfrac{1}{2}\ln 2p + \tfrac{1}{2}\ln 2(1-p), \end{aligned}$$

(where (a) and (b) follow from (21) and (22), respectively) whence

$$\begin{aligned} \lim_{n\to\infty} \sqrt[n]{\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y)} &= \exp\left(\tfrac{1}{2}\ln(2p) + \tfrac{1}{2}\ln(2(1-p))\right) \\ &= 2\sqrt{p(1-p)}. \end{aligned} \tag{23}$$

On the other hand,

$$\Phi = \sum_{i=1}^{n}(p_i - \tfrac{1}{2})\log\frac{p_i}{1-p_i} = n(p-\tfrac{1}{2})\log\frac{p}{1-p},$$

and hence

$$\sqrt[n]{U} = [(1-p)/p]^{(p-\frac{1}{2})/2}.$$

The tightness of the upper bound follows from

$$F(p) := \frac{2\sqrt{p(1-p)}}{[(1-p)/p]^{(p-\frac{1}{2})/2}} \xrightarrow[p\to 1/2]{} 1,$$

which is easily verified since $F(\frac{1}{2}) = 1$.

Lower bound. For the lower bound, consider a single expert with competence $p_1 = p > \frac{1}{2}$. Thus, $\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) = 1 - p$ and $L \asymp \exp(-2\Phi) = [(1-p)/p]^{2p-1}$. Again, it is easily verified that

$$\frac{[(1-p)/p]^{2p-1}}{1-p} \quad \underset{p\to 1}{\longrightarrow} \quad 1,$$

and so the lower bound is also tight.

We conclude that the committee profile Φ is not sufficiently sensitive an indicator to close the gap between the two bounds entirely.

Remark 6 *In the special case of identical experts, with $p_1 = \ldots = p_n = p$, the Chernoff-Stein lemma (Cover and Thomas, 2006) gives the best asymptotic exponent for one-sided (i.e., type I or type II) errors, while* Chernoff information *corresponds to the optimal exponent for the overall probability of error. As seen from (23), the latter is given by $\frac{1}{2}\ln(2p) + \frac{1}{2}\ln(2(1-p))$ in this case.*

In contradistinction, our bounds in Theorem 1 hold for non-identical experts and are dimension-free.

3.4 Additional bounds

An anonymous referee has pointed out that

$$\Phi = \tfrac{1}{2}D(P||Q) = \tfrac{1}{2}D(Q||P), \tag{24}$$

where P is the distribution of $\boldsymbol{\eta} \in \{\pm 1\}^n$ defined in (13), Q is the "antipodal" distribution of $-\boldsymbol{\eta}$, and $D(P||Q)$ is the Kullback-Leibler divergence, defined by

$$D(P||Q) = \sum_{\mathbf{x}\in\{\pm 1\}^n} P(\mathbf{x}) \ln \frac{P(\mathbf{x})}{Q(\mathbf{x})}.$$

This leads to an improved lower bound for $\Phi \lesssim 0.992$, as follows. By Lemma 2, we have

$$\begin{aligned} \mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) &= \tfrac{1}{2} \sum_{\boldsymbol{\eta}\in\{\pm 1\}^n} \min\{P(\boldsymbol{\eta}), Q(\boldsymbol{\eta})\} \\ &= \tfrac{1}{2}\left(1 - \tfrac{1}{2}\left\|P - Q\right\|_1\right), \end{aligned} \tag{25}$$

where the second identity follows from a well-known minorization characterization of the total variation distance (see, e.g., Kontorovich (2007, Lemma 2.2.2)). A bound relating the total variation distance and Kullback-Leibler divergence is known as Pinsker's inequality, and states that

$$\left\|P - Q\right\|_1 \leq \sqrt{2D(P||Q)} \tag{26}$$

holds for all distributions P, Q (see Berend et al. (2014) for historical background and a "reversed" direction of (26)). Combining (24), (25), and (26), we obtain

$$\mathbb{P}(f^{\text{OPT}}(\mathbf{X}) \neq Y) \geq \tfrac{1}{2}\left(1 - \sqrt{\Phi}\right),$$

which, for small Φ, is far superior to Theorem 1(ii) (but is vacuous for $\Phi \geq 1$).

The identity in (25) may also be used to sharpen the upper bound in Theorem 1(i) for small Φ. Invoking Even-Dar et al. (2007, Lemma 3.10), we have

$$D(P||Q) \leq \|P - Q\|_1 \log \left(\min_{\mathbf{x} \in \{\pm 1\}^n} P(\mathbf{x}) \right)^{-1}. \tag{27}$$

Let us suppose for concreteness that all of the experts are identical with $p_i = \frac{1}{2} + \gamma$ for $\gamma \in (0, \frac{1}{2})$, $i \in [n]$. Then

$$\Phi = n\gamma \log \frac{1/2 + \gamma}{1/2 - \gamma}$$

and

$$\log \left(\min_{\mathbf{x} \in \{\pm 1\}^n} P(\mathbf{x}) \right)^{-1} = n \log \frac{1}{1/2 - \gamma} =: \Gamma,$$

which, combined with (24, 25, 27) yields

$$\begin{aligned} \mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) \neq Y) &\leq \tfrac{1}{2} \left(1 - \frac{\Phi}{\Gamma} \right) \\ &= \tfrac{1}{2} \left(1 - \gamma + \gamma \frac{\log(1/2 + \gamma)}{\log(1/2 - \gamma)} \right). \end{aligned} \tag{28}$$

Thus, for $0 < \gamma < \frac{1}{2}$ and

$$n < \frac{2}{\gamma} \left(\log \frac{1/2 - \gamma}{1/2 + \gamma} \right) \log \left(\frac{1 - \gamma}{2} + \frac{\gamma}{2} \cdot \frac{\log(1/2 + \gamma)}{\log(1/2 - \gamma)} \right),$$

(28) is sharper than Theorem 1(i).

4. Unknown Competences: Frequentist Approach

Our goal in this section is to obtain, insofar as possible, analogues of Theorem 1 for unknown expert competences. When the p_is are unknown, they must be estimated empirically before any useful weighted majority vote can be applied. There are various ways to model partial knowledge of expert competences (Baharad et al., 2011, 2012). Perhaps the simplest scenario for estimating the p_is is to assume that the i^{th} expert has been queried independently m_i times, out of which he gave the correct prediction k_i times. Taking the $\{m_i\}$ to be fixed, define the *committee profile* by $\mathbf{k} = (k_1, \ldots, k_n)$; this is the aggregate of the agent's empirical knowledge of the experts' performance. An *empirical decision rule* $\hat{f} : (\mathbf{x}, \mathbf{k}) \mapsto \{\pm 1\}$ makes a final decision based on the expert inputs $\mathbf{x}$ together with the committee profile. Analogously to (1), the probability of a mistake is

$$\mathbb{P}(\hat{f}(\mathbf{X}, \mathbf{K}) \neq Y). \tag{29}$$

Note that now the committee profile is an additional source of randomness. Here we run into our first difficulty: unlike the probability in (1), which is minimized by the naive Bayes

decision rule, the agent cannot formulate an optimal decision rule $\hat{f}$ in advance without knowing the p_is. This is because no decision rule is optimal uniformly over the range of possible p_is. Our approach will be to consider weighted majority decision rules of the form

$$\hat{f}(\mathbf{x}, \mathbf{k}) = \operatorname{sign}\left(\sum_{i=1}^{n} \hat{w}(k_i) x_i\right) \tag{30}$$

and to analyze their consistency properties under two different regimes: low-confidence and high-confidence. These refer to the confidence intervals of the frequentist estimate of p_i, given by

$$\hat{p}_i = \frac{k_i}{m_i}. \tag{31}$$

4.1 Low-confidence regime

In the low-confidence regime, the sample sizes m_i may be as small as 1, and we define[3]

$$\hat{w}(k_i) = \hat{w}_i^{\mathrm{LC}} := \hat{p}_i - \tfrac{1}{2}, \qquad i \in [n], \tag{32}$$

which induces the empirical decision rule $\hat{f}^{\mathrm{LC}}$. It remains to analyze $\hat{f}^{\mathrm{LC}}$'s probability of error. Recall the definition of ξ_i from (5) and observe that

$$\mathbb{E}\left[\hat{w}_i^{\mathrm{LC}} \xi_i\right] = \mathbb{E}[(\hat{p}_i - \tfrac{1}{2})\xi_i] = (p_i - \tfrac{1}{2}) p_i, \tag{33}$$

since $\hat{p}_i$ and ξ_i are independent. As in (6), the probability of error (29) is

$$\mathbb{P}\left(\sum_{i=1}^{n} \hat{w}_i^{\mathrm{LC}} \xi_i \le \frac{1}{2} \sum_{i=1}^{n} \hat{w}_i^{\mathrm{LC}}\right) = \mathbb{P}\left(\sum_{i=1}^{n} Z_i \le 0\right), \tag{34}$$

where $Z_i = \hat{w}_i^{\mathrm{LC}}(\xi_i - \frac{1}{2})$. Now the $\{Z_i\}$ are independent random variables, $\mathbb{E}Z_i = (p_i - \frac{1}{2})^2$ (by (33)), and each Z_i takes values in an interval of length $\frac{1}{2}$. Hence, the standard Hoeffding bound applies:

$$\mathbb{P}(\hat{f}^{\mathrm{LC}}(\mathbf{X}, \mathbf{K}) \neq Y) \le \exp\left[-\frac{8}{n}\left(\sum_{i=1}^{n} (p_i - \tfrac{1}{2})^2\right)^2\right]. \tag{35}$$

We summarize these calculations in

Theorem 7 *A sufficient condition*[4] *for* $\mathbb{P}(\hat{f}^{\mathrm{LC}}(\mathbf{X}, \mathbf{K}) \neq Y) \to 0$ *is*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (p_i - \tfrac{1}{2})^2 \to \infty.$$

3. For $m_i \min\{p_i, q_i\} \ll 1$, the estimated competences $\hat{p}_i$ may well take values in $\{0, 1\}$, in which case $\log(\hat{p}_i/\hat{q}_i) = \pm\infty$. The rule in (32) is essentially a first-order Taylor approximation to $w(\cdot)$ about $p = \frac{1}{2}$.

4. Formally, we have an infinite sequence of experts with competences $\{p_i : i \in \mathbb{N}\}$, with a corresponding sequence of trials with sizes $\{m_i\}$ and outcomes $K_i \sim \mathrm{Bin}(m_i, p_i)$, in addition to the expert votes $X_i \sim Y\,[2 \cdot \mathrm{Bernoulli}(p_i) - 1]$. An empirical decision rule f_n (more precisely, a sequence of rules) is said to be *consistent* if

$$\lim_{n \to \infty} \mathbb{P}(f_n(\mathbf{X}, \mathbf{K}) \neq Y) = 0.$$

Several remarks are in order. First, notice that the error bound in (35) is stated in terms of the unknown $\{p_i\}$, providing the agent with large-committee asymptotics but giving no finitary information; this limitation is inherent in the low-confidence regime. Secondly, the condition in Theorem 7 is considerably more restrictive than the consistency condition $\Phi \to \infty$ implicit in Theorem 1. Indeed, the empirical decision rule $\hat{f}^{\mathrm{LC}}$ is incapable of exploiting a single highly competent expert in the way that f^{OPT} from (2) does. Our analysis could be sharpened somewhat for moderate sample sizes $\{m_i\}$ by using Bernstein's inequality to take advantage of the low variance of the $\hat{p}_i$s. For sufficiently large sample sizes, however, the high-confidence regime (discussed below) begins to take over. Finally, there is one sense in which this case is "easier" to analyze than that of known $\{p_i\}$: since the summands in (34) are bounded, Hoeffding's inequality gives nontrivial results and there is no need for more advanced tools such as the Kearns-Saul inequality (9) (which is actually inapplicable in this case).

4.2 High-confidence regime

In the high-confidence regime, each estimated competence $\hat{p}_i$ is close to the true value p_i with high probability. To formalize this, fix some $0 < \delta < 1$, $0 < \varepsilon \leq 5$, and put

$$q_i = 1 - p_i, \;\; \hat{q}_i = 1 - \hat{p}_i.$$

We will set the empirical weights according to the "plug-in" naive Bayes rule

$$\hat{w}_i^{\mathrm{HC}} := \log \frac{\hat{p}_i}{\hat{q}_i}, \qquad i \in [n], \tag{36}$$

which induces the empirical decision rule $\hat{f}^{\mathrm{HC}}$ and raises immediate concerns about $\hat{w}_i^{\mathrm{HC}} = \pm\infty$. We give two kinds of bounds on $\mathbb{P}(\hat{f}^{\mathrm{HC}} \neq Y)$: nonadaptive and adaptive. In the nonadaptive analysis, we show that for $m_i \min\{p_i, q_i\} \gg 1$, with high probability $|w_i - \hat{w}_i^{\mathrm{HC}}| \ll 1$, and thus a "perturbed" version of Theorem 1(i) holds (and in particular, w_i^{HC} will be finite with high probability). In the adaptive analysis, we allow $\hat{w}_i^{\mathrm{HC}}$ to take on infinite values[5] and show (perhaps surprisingly) that this decision rule still admits reasonable error estimates.

Nonadaptive analysis. In this section, $\varepsilon, \tilde{\varepsilon} > 0$ are related by $\varepsilon = 2\tilde{\varepsilon} + 4\tilde{\varepsilon}^2$ or, equivalently,

$$\tilde{\varepsilon} = \frac{\sqrt{4\varepsilon + 1} - 1}{4}. \tag{37}$$

Lemma 8 *If* $0 < \tilde{\varepsilon} < 1$ *and*

$$\tilde{\varepsilon}^2 m_i p_i \;\; \geq \;\; 3\log(2n/\delta), \qquad i \in [n], \tag{38}$$

then

$$\mathbb{P}\left(\exists i \in [n] : \frac{\hat{p}_i}{p_i} \notin (1 - \tilde{\varepsilon}, 1 + \tilde{\varepsilon})\right) \leq \delta.$$

5. When the decision rule is faced with evaluating sums involving $\infty - \infty$, we automatically count this as an error.

Proof The multiplicative Chernoff bound yields

$$\mathbb{P}\left(\hat{p}_i < (1-\tilde{\varepsilon})p_i\right) \leq e^{-\tilde{\varepsilon}^2 m_i p_i/2}$$

and

$$\mathbb{P}\left(\hat{p}_i > (1+\tilde{\varepsilon})p_i\right) \leq e^{-\tilde{\varepsilon}^2 m_i p_i/3}.$$

Hence,

$$\mathbb{P}\left(\frac{\hat{p}_i}{p_i} \notin (1-\tilde{\varepsilon}, 1+\tilde{\varepsilon})\right) \quad \leq \quad 2e^{-\tilde{\varepsilon}^2 m_i p_i/3}.$$

The claim follows from (38) and the union bound. ■

Lemma 9 *Let $\delta \in (0,1)$, $\varepsilon \in (0,5)$, and w_i be the naive Bayes weight (3). If*

$$1-\tilde{\varepsilon} \leq \frac{\hat{p}_i}{p_i}, \frac{\hat{q}_i}{q_i} \leq 1+\tilde{\varepsilon}$$

then

$$|w_i - \hat{w}_i^{\mathrm{HC}}| \leq \varepsilon.$$

Proof We have

$$\begin{aligned} |w_i - \hat{w}_i^{\mathrm{HC}}| &= \left|\log\frac{p_i}{q_i} - \log\frac{\hat{p}_i}{\hat{q}_i}\right| \\ &= \left|\log\frac{p_i}{\hat{p}_i} + \log\frac{\hat{q}_i}{q_i}\right| \\ &= \left|\log\frac{p_i}{\hat{p}_i}\right| + \left|\log\frac{\hat{q}_i}{q_i}\right|. \end{aligned}$$

Now[6]

$$\begin{aligned} [\log(1-\tilde{\varepsilon}), \log(1+\tilde{\varepsilon})] &\subseteq [-\tilde{\varepsilon} - 2\tilde{\varepsilon}^2, \tilde{\varepsilon}] \\ &\subseteq [-\tfrac{1}{2}\varepsilon, \tfrac{1}{2}\varepsilon], \end{aligned}$$

whence

$$\left|\log\frac{p_i}{\hat{p}_i}\right| + \left|\log\frac{\hat{q}_i}{q_i}\right| \leq \varepsilon.$$

■

6. The first containment requires $\log(1-x) \geq -x - 2x^2$, which holds (not exclusively) on $(0, 0.9)$. The restriction $\varepsilon \leq 5$ ensures that $\tilde{\varepsilon}$ is in this range.

Corollary 10 *If*

$$\tilde{\varepsilon}^2 m_i \min\{p_i, q_i\} \;\geq\; 3\log(4n/\delta), \qquad i \in [n],$$

then

$$\mathbb{P}\left(\max_{i\in[n]} |w_i - \hat{w}_i^{\mathrm{HC}}| > \varepsilon\right) \;\leq\; \delta.$$

Proof An immediate consequence of applying Lemma 8 to p_i and q_i with the union bound. ■

To state the next result, let us arrange the plug-in weights (36) as a vector $\hat{\mathbf{w}}^{\mathrm{HC}} \in \mathbb{R}^n$, as was done with $\mathbf{w}$ and $\boldsymbol{\eta}$ from Section 3.1. The corresponding weighted majority rule $\hat{f}^{\mathrm{HC}}$ yields an error precisely when

$$\hat{\mathbf{w}}^{\mathrm{HC}} \cdot \boldsymbol{\eta} \leq 0$$

(cf. (14)). Our nonadaptive approach culminates in the following result.

Theorem 11 *Let* $0 < \delta < 1$ *and* $0 < \varepsilon < \min\{5, 2\Phi/n\}$. *If*

$$m_i \min\{p_i, q_i\} \geq 3\left(\frac{\sqrt{4\varepsilon+1}-1}{4}\right)^{-2} \log\frac{4n}{\delta}, \qquad i \in [n], \tag{39}$$

then

$$\mathbb{P}\left(\hat{f}^{\mathrm{HC}}(\mathbf{X}, \mathbf{K}) \neq Y\right) \;\leq\; \delta + \exp\left[-\frac{(2\Phi - \varepsilon n)^2}{8\Phi}\right]. \tag{40}$$

Remark 12 *For fixed* $\{p_i\}$ *different from* 0 *or* 1 *and* $\min_{i\in[n]} m_i \to \infty$, *we may take* δ *and* ε *arbitrarily small — and in this limiting case, the bound of Theorem 1(i) is recovered.*

Proof Suppose that Z, $\hat{Z}$, and U are real numbers satisfying

$$\left|Z - \hat{Z}\right| \leq U.$$

Then

$$\forall t > 0, \quad (\hat{Z} \leq 0) \implies (U > t) \vee (Z \leq t). \tag{41}$$

Indeed, if both $U \leq t$ and $Z > t$, then $\hat{Z}$ and Z are within a distance t of each other, but $Z > t$ and so $\hat{Z}$ must be greater than 0.

Observe also that $\|\boldsymbol{\eta}\|_\infty = 1$, and thus a simple application of Hölder's inequality yields

$$\begin{aligned} |\mathbf{w}\cdot\boldsymbol{\eta} - \hat{\mathbf{w}}^{\mathrm{HC}}\cdot\boldsymbol{\eta}| &= |(\mathbf{w} - \hat{\mathbf{w}}^{\mathrm{HC}})\cdot\boldsymbol{\eta}| \\ &\leq \sum_{i=1}^{n} |w_i - w_i^{\mathrm{HC}}| = \|\mathbf{w} - \hat{\mathbf{w}}^{\mathrm{HC}}\|_1 . \end{aligned}$$

Invoking (41) with $Z = \mathbf{w}$, $\hat{Z} = \hat{\mathbf{w}}^{\mathrm{HC}}$, and $t = \varepsilon n$, we obtain

$$\begin{aligned}\mathbb{P}\left(\hat{\mathbf{w}}^{\mathrm{HC}} \cdot \boldsymbol{\eta} \leq 0\right) &\leq \mathbb{P}\left(\{\|\mathbf{w} - \hat{\mathbf{w}}^{\mathrm{HC}}\|_1 > \varepsilon n\} \cup \{\mathbf{w} \cdot \boldsymbol{\eta} \leq \varepsilon n\}\right) \\ &\leq \mathbb{P}(\|\mathbf{w} - \hat{\mathbf{w}}^{\mathrm{HC}}\|_1 > \varepsilon n) + \mathbb{P}(\mathbf{w} \cdot \boldsymbol{\eta} \leq \varepsilon n).\end{aligned}$$

Corollary 10 upper-bounds the first term on the right-hand side by δ. The second term is estimated by replacing Φ by $\Phi - \varepsilon n$ in (10) and repeating the argument following that formula. ■

Adaptive analysis. Theorem 11 has the drawback of being *nonadaptive*, in that its assumptions (39) and conclusions (40) depend on the unknown $\{p_i\}$ and hence cannot be evaluated by the agent (the bound in Display 35 is also nonadaptive). In the *adaptive* approach, all results are stated in terms of empirically observed quantities:

Theorem 13 *Choose any*

$$\delta \geq \sum_{i=1}^{n} \frac{1}{\sqrt{m_i}}$$

and let R be the event

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\hat{p}_i - \tfrac{1}{2})\hat{w}_i^{\mathrm{HC}}\right) \leq \frac{\delta}{2}. \tag{42}$$

Then

$$\mathbb{P}\left(R \cap \left\{\hat{f}^{\mathrm{HC}}(\mathbf{X}, \mathbf{K}) \neq Y\right\}\right) \leq \delta.$$

Remark 14 *Our interpretation for Theorem 13 is as follows. The agent observes the committee profile $\mathbf{K}$, which determines the $\{\hat{p}_i, \hat{w}_i^{\mathrm{HC}}\}$, and then checks whether the event R has occurred. If not, the adaptive agent refrains from making a decision (and may choose to fall back on the low-confidence approach described previously). If R* does *hold, however, the agent predicts Y according to $\hat{f}^{\mathrm{HC}}$. The event R will tend to occur when the estimated $\hat{p}_i$s are "favorable" in the sense of inducing a large empirical committee profile. When this fails to happen (i.e., many of the $\hat{p}_i$ are close to $\frac{1}{2}$), R will be a rare event. However, in this case little is lost by refraining from a high-confidence decision and defaulting to a low-confidence one, since near $\frac{1}{2}$, the two decision procedures are very similar.*

As explained above, there does not exist a nontrivial a priori upper bound on $\mathbb{P}(\hat{f}^{\mathrm{HC}}(\mathbf{X}, \mathbf{K}) \neq Y)$ independent of any knowledge of the p_is. Instead, Theorem 13 bounds the probability of the agent being "fooled" by an unrepresentative committee profile.[7] *Note that we have done nothing to prevent $\hat{w}_i^{\mathrm{HC}} = \pm\infty$, and this may indeed happen. Intuitively, there are two reasons for infinite $\hat{w}_i^{\mathrm{HC}}$: (a) noisy $\hat{p}_i$ due to m_i being too small, or (b) the i^{th} expert is actually highly (in)competent, which causes $\hat{p}_i \in \{0, 1\}$ to be likely even for large m_i. The $1/\sqrt{m_i}$ term in the bound insures against case (a), while in case (b), choosing infinite $\hat{w}_i^{\mathrm{HC}}$ causes no harm (as we show in the proof).*

7. These adaptive bounds are similar in spirit to *empirical Bernstein* methods, (Audibert et al., 2007; Mnih et al., 2008; Maurer and Pontil, 2009), where the agent's confidence depends on the empirical variance.

Proof We will write the probability and expectation operators with subscripts (such as $\mathbf{K}$) to indicate the random variable(s) being summed over. Thus,

$$\begin{aligned}\mathbb{P}_{\mathbf{K},\mathbf{X},Y}\left(R\cap\left\{\hat{f}^{\mathrm{HC}}(\mathbf{X},\mathbf{K})\neq Y\right\}\right) &= \mathbb{P}_{\mathbf{K},\boldsymbol{\eta}}\left(R\cap\{\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\boldsymbol{\eta}\leq 0\}\right)\\ &= \mathbb{E}_{\mathbf{K}}\left[\mathbb{1}_R\cdot\mathbb{P}_{\boldsymbol{\eta}}\left(\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\boldsymbol{\eta}\leq 0\,|\,\mathbf{K}\right)\right].\end{aligned} \tag{43}$$

Recall that the random variable $\boldsymbol{\eta}\in\{\pm 1\}^n$, with probability mass function

$$P(\boldsymbol{\eta})=\prod_{i:\eta_i=1}p_i\prod_{i:\eta_i=-1}q_i,$$

is independent of $\mathbf{K}$, and hence

$$\mathbb{P}_{\boldsymbol{\eta}}\left(\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\boldsymbol{\eta}\leq 0\,|\,\mathbf{K}\right)=\mathbb{P}_{\boldsymbol{\eta}}\left(\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\boldsymbol{\eta}\leq 0\right). \tag{44}$$

Define the random variable $\hat{\boldsymbol{\eta}}\in\{\pm 1\}^n$ (conditioned on $\mathbf{K}$) by the probability mass function

$$P(\hat{\boldsymbol{\eta}})=\prod_{i:\eta_i=1}\hat{p}_i\prod_{i:\eta_i=-1}\hat{q}_i,$$

and the set $A\subseteq\{\pm 1\}^n$ by $A=\{\mathbf{x}:\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\mathbf{x}\leq 0\}$. Now

$$\begin{aligned}\left|\mathbb{P}_{\boldsymbol{\eta}}\left(\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\boldsymbol{\eta}\leq 0\right)-\mathbb{P}_{\hat{\boldsymbol{\eta}}}\left(\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\hat{\boldsymbol{\eta}}\leq 0\right)\right| &= \left|\mathbb{P}_{\boldsymbol{\eta}}(A)-\mathbb{P}_{\hat{\boldsymbol{\eta}}}(A)\right|\\ &\leq \max_{A\subseteq\{\pm 1\}^n}\left|\mathbb{P}_{\boldsymbol{\eta}}(A)-\mathbb{P}_{\hat{\boldsymbol{\eta}}}(A)\right|\\ &= \left\|\mathbb{P}_{\boldsymbol{\eta}}-\mathbb{P}_{\hat{\boldsymbol{\eta}}}\right\|_{\mathrm{TV}}\\ &\leq \sum_{i=1}^n|p_i-\hat{p}_i|=:M,\end{aligned}$$

where the last inequality follows from a standard tensorization property of the total variation norm $\|\cdot\|_{\mathrm{TV}}$, see e.g. (Kontorovich, 2012, Lemma 2.2). By Theorem 1(i), we have

$$\mathbb{P}_{\hat{\boldsymbol{\eta}}}\left(\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\hat{\boldsymbol{\eta}}\leq 0\right)\leq\exp\left(-\tfrac{1}{2}\sum_{i=1}^n(\hat{p}_i-\tfrac{1}{2})\hat{w}_i^{\mathrm{HC}}\right),$$

and hence

$$\mathbb{P}_{\boldsymbol{\eta}}\left(\hat{\mathbf{w}}^{\mathrm{HC}}\cdot\boldsymbol{\eta}\leq 0\right)\leq M+\exp\left(-\tfrac{1}{2}\sum_{i=1}^n(\hat{p}_i-\tfrac{1}{2})\hat{w}_i^{\mathrm{HC}}\right).$$

Invoking (44), we substitute the right-hand side above into (43) to obtain

$$\begin{aligned}\mathbb{P}_{\mathbf{K},\mathbf{X},Y}\left(R\cap\left\{\hat{f}^{\mathrm{HC}}(\mathbf{X},\mathbf{K})\neq Y\right\}\right) &\leq \mathbb{E}_{\mathbf{K}}\left[\mathbb{1}_R\cdot\left(M+\exp\left(-\tfrac{1}{2}\sum_{i=1}^n(\hat{p}_i-\tfrac{1}{2})\hat{w}_i^{\mathrm{HC}}\right)\right)\right]\\ &\leq \mathbb{E}_{\mathbf{K}}[M]+\mathbb{E}_{\mathbf{K}}\left[\mathbb{1}_R\exp\left(-\tfrac{1}{2}\sum_{i=1}^n(\hat{p}_i-\tfrac{1}{2})\hat{w}_i^{\mathrm{HC}}\right)\right].\end{aligned}$$

By the definition of R, the second term on the last right-hand side is upper-bounded by $\delta/2$. To bound M, we invoke a simple mean absolute deviation estimate (cf. Berend and Kontorovich, 2013a):

$$\mathbb{E}_{\mathbf{K}} |p_i - \hat{p}_i| \;\leq\; \sqrt{\frac{p_i(1-p_i)}{m_i}} \;\leq\; \frac{1}{2\sqrt{m_i}},$$

which finishes the proof. ■

Remark 15 *Actually, the proof shows that we may take a smaller δ, but with a more complex dependence on $\{m_i\}$, which simplifies to $2[1-(1-(2\sqrt{m})^{-1})^n]$ for $m_i \equiv m$. This improvement is achieved via a refinement of the bound $\left\|\mathbb{P}_{\boldsymbol{\eta}} - \mathbb{P}_{\hat{\boldsymbol{\eta}}}\right\|_{\mathrm{TV}} \leq \sum_{i=1}^n |p_i - \hat{p}_i|$ to $\left\|\mathbb{P}_{\boldsymbol{\eta}} - \mathbb{P}_{\hat{\boldsymbol{\eta}}}\right\|_{\mathrm{TV}} \leq \alpha\left(\{|p_i - \hat{p}_i| : i \in [n]\}\right)$, where $\alpha(\cdot)$ is the function defined in Kontorovich (2012, Lemma 4.2).*

Open problem. As argued in Remark 12, the nonadaptive agent achieves the asymptotically optimal rate of Theorem 1(i) in the large-sample limit. Does an analogous claim hold true for the adaptive agent? Can the dependence on $\{m_i\}$ in Theorem 13 be improved, perhaps through a better choice of $\hat{\mathbf{w}}^{\mathrm{HC}}$?

5. Unknown Competences: Bayesian Approach

A shortcoming of Theorem 13 is that, when condition R fails, the agent is left with no estimate of the error probability. An alternative (and in some sense cleaner) approach to handling unknown expert competences p_i is to assume a known prior distribution over the competence levels p_i. The natural choice of prior for a Bernoulli parameter is the Beta distribution, namely

$$p_i \sim \mathrm{Beta}(\alpha_i, \beta_i)$$

with density

$$\frac{p_i^{\alpha_i - 1} q_i^{\beta_i - 1}}{B(\alpha_i, \beta_i)}, \qquad \alpha_i, \beta_i > 0,$$

where $q_i = 1 - p_i$ and $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$. Our full probabilistic model is as follows. First, "nature" chooses the true state of the world Y according to $Y \sim \mathrm{Bernoulli}(\frac{1}{2})$, and each of the n expert competences p_i is drawn independently from $\mathrm{Beta}(\alpha_i, \beta_i)$ with known parameters α_i, β_i. Then the i^{th} expert, $i \in [n]$, is queried (on independent instances) m_i times, with $K_i \sim \mathrm{Bin}(m_i, p_i)$ correct predictions and $m_i - K_i$ incorrect ones. As before, $\mathbf{K} = (K_1, \ldots, K_n)$ is the (random) committee profile. Additionally, $\mathbf{X} = (X_1, \ldots, X_n)$ is the random voting profile, where $X_i \sim Y\,[2 \cdot \mathrm{Bernoulli}(p_i) - 1]$, independent of the other random variables. Absent direct knowledge of the p_is, the agent relies on an empirical decision rule $\hat{f} : (\mathbf{x}, \mathbf{k}) \mapsto \{\pm 1\}$ to produce a final decision based on the expert inputs $\mathbf{x}$ together with the committee profile $\mathbf{k}$. A decision rule $\hat{f}^{\mathrm{Ba}}$ is *Bayes-optimal* if it minimizes

$$\mathbb{P}(\hat{f}(\mathbf{X}, \mathbf{K}) \neq Y), \tag{45}$$

which is formally identical to (29) but semantically there is a difference: the probability in (45) is over the p_i in addition to $(\mathbf{X}, Y, \mathbf{K})$. Unlike the frequentist approach, where no optimal empirical decision rule was possible, the Bayesian approach readily admits one:

Theorem 16 *The decision rule*

$$\hat{f}^{\mathrm{Ba}}(\mathbf{x}, \mathbf{k}) = \operatorname{sign}\left(\sum_{i=1}^{n} \hat{w}_i^{\mathrm{Ba}} x_i\right), \tag{46}$$

where

$$\hat{w}_i^{\mathrm{Ba}} = \log \frac{\alpha_i + k_i}{\beta_i + m_i - k_i}, \tag{47}$$

minimizes the probability in (45) over all empirical decision rules.

Remark 17 *For $0 < p_i < 1$, we have*

$$\hat{w}_i^{\mathrm{Ba}} \underset{m_i \to \infty}{\longrightarrow} w_i, \qquad i \in [n],$$

almost surely, both in the frequentist and the Bayesian interpretations.

Proof Denote

$$M_n \;=\; \{0, \ldots, m_1\} \times \{0, \ldots, m_2\} \times \ldots \times \{0, \ldots, m_n\}$$

and let $f : \{\pm 1\}^n \times M_n \to \{\pm 1\}$ be an arbitrary empirical decision rule. Then

$$\mathbb{P}(f(\mathbf{X}, \mathbf{K}) \neq Y) \;=\; \sum_{\mathbf{x} \in \{\pm 1\}^n,\ \mathbf{k} \in M_n} \mathbb{P}(\mathbf{X} = \mathbf{x}, \mathbf{K} = \mathbf{k}) \cdot \mathbb{P}(f(\mathbf{X}, \mathbf{K}) \neq Y \,|\, \mathbf{X} = \mathbf{x}, \mathbf{K} = \mathbf{k}).$$

Observe that the quantity $\mathbb{P}(Y = y \,|\, \mathbf{X} = \mathbf{x}, \mathbf{K} = \mathbf{k})$ is completely determined by y, $\mathbf{x}$, $\mathbf{k}$, and the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$, and denote this functional dependence by

$$\mathbb{P}(Y = y \,|\, \mathbf{X} = \mathbf{x}, \mathbf{K} = \mathbf{k}) \;=:\; G_{\boldsymbol{\alpha},\boldsymbol{\beta}}(y, \mathbf{x}, \mathbf{k}).$$

Then clearly, the optimal empirical decision rule is

$$f^*_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{x}, \mathbf{k}) \;=\; \begin{cases} +1, & G_{\boldsymbol{\alpha},\boldsymbol{\beta}}(+1, \mathbf{x}, \mathbf{k}) \geq G_{\boldsymbol{\alpha},\boldsymbol{\beta}}(-1, \mathbf{x}, \mathbf{k}), \\ -1, & G_{\boldsymbol{\alpha},\boldsymbol{\beta}}(+1, \mathbf{x}, \mathbf{k}) < G_{\boldsymbol{\alpha},\boldsymbol{\beta}}(-1, \mathbf{x}, \mathbf{k}), \end{cases}$$

and a decision rule $f_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ is optimal if and only if

$$\mathbb{P}(f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}, \mathbf{K}) = Y \,|\, \mathbf{X} = \mathbf{x}, \mathbf{K} = \mathbf{k}) \;\geq\; \mathbb{P}(f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}, \mathbf{K}) \neq Y \,|\, \mathbf{X} = \mathbf{x}, \mathbf{K} = \mathbf{k}) \tag{48}$$

for all $\mathbf{x}, \mathbf{k}, \boldsymbol{\alpha}, \boldsymbol{\beta}$. Invoking Bayes' formula, we may rewrite the optimality criterion in (48) in the form

$$\mathbb{P}(f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}, \mathbf{K}) = Y, \mathbf{X} = \mathbf{x}, \mathbf{K} = \mathbf{k}) \;\geq\; \mathbb{P}(f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}, \mathbf{K}) \neq Y, \mathbf{X} = \mathbf{x}, \mathbf{K} = \mathbf{k}). \tag{49}$$

For given $\mathbf{x} \in \{\pm 1\}^n$ and $\mathbf{k} \in M_n$, let $I_+(\mathbf{x})$ be the set of YES votes

$$I_+(\mathbf{x}) = \{i \in [n] : x_i = +1\}$$

and $I_-(\mathbf{x}) = [n] \setminus I_+(\mathbf{x})$ the set of NO votes. Let us fix some $A \subseteq [n]$, $B = [n] \setminus A$ and compute

$$\begin{aligned}
&\mathbb{P}(Y = +1, I_+(\mathbf{X}) = A, I_-(\mathbf{X}) = B, \mathbf{k} = \mathbf{K}) \\
&= \prod_{i=1}^n \int_0^1 \frac{p_i^{\alpha_i - 1} q_i^{\beta_i - 1}}{B(\alpha_i, \beta_i)} \binom{m_i}{k_i} p_i^{k_i} q_i^{m_i - k_i} p_i^{\mathbb{1}_{\{i \in A\}}} q_i^{\mathbb{1}_{\{i \in B\}}} dp_i \\
&= \prod_{i=1}^n \frac{\binom{m_i}{k_i}}{B(\alpha_i, \beta_i)} \int_0^1 p_i^{\alpha_i + k_i - 1 + \mathbb{1}_{\{i \in A\}}} q_i^{\beta_i + m_i - k_i - 1 + \mathbb{1}_{\{i \in B\}}} dp_i \\
&= \prod_{i=1}^n \frac{\binom{m_i}{k_i} B(\alpha_i + k_i + \mathbb{1}_{\{i \in A\}}, \beta_i + m_i - k_i + \mathbb{1}_{\{i \in B\}})}{B(\alpha_i, \beta_i)}. \qquad (50)
\end{aligned}$$

Analogously,

$$\begin{aligned}
&\mathbb{P}(Y = -1, I_+(\mathbf{X}) = A, I_-(\mathbf{X}) = B, \mathbf{k} = \mathbf{K}) \\
&= \prod_{i=1}^n \frac{\binom{m_i}{k_i} B(\alpha_i + k_i + \mathbb{1}_{\{i \in B\}}, \beta_i + m_i - k_i + \mathbb{1}_{\{i \in A\}})}{B(\alpha_i, \beta_i)}. \qquad (51)
\end{aligned}$$

Let us use the shorthand $P(+1, A, B, \mathbf{k})$ and $P(-1, A, B, \mathbf{k})$ for the joint probabilities in the last two displays, along with their corresponding conditionals $P(\pm 1 \,|\, A, B, \mathbf{k})$. Obviously,

$$P(1|A, B, \mathbf{k}) > P(-1|A, B, \mathbf{k}) \iff P(1, A, B, \mathbf{k}) > P(-1, A, B, \mathbf{k}),$$

which occurs precisely if

$$\prod_{i=1}^n B(\alpha_i + k_i + \mathbb{1}_{\{i \in A\}}, \beta_i + m_i - k_i + \mathbb{1}_{\{i \in B\}}) > \prod_{i=1}^n B(\alpha_i + k_i + \mathbb{1}_{\{i \in B\}}, \beta_i + m_i - k_i + \mathbb{1}_{\{i \in A\}}), \qquad (52)$$

as the other factors in (50) and (51) cancel out. Now $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$ and

$$\begin{aligned}
\Gamma(\alpha_i + k_i + \mathbb{1}_{\{i \in A\}} + \beta_i + m_i - k_i + \mathbb{1}_{\{i \in B\}}) &= \Gamma(\alpha_i + k_i + \mathbb{1}_{\{i \in B\}} + \beta_i + m_i - k_i + \mathbb{1}_{\{i \in A\}}) \\
&= \Gamma(\alpha_i + \beta_i + m_i + 1),
\end{aligned}$$

and thus both sides of (52) share a common factor of

$$\left(\prod_{i=1}^n \Gamma(\alpha_i + \beta_i + m_i + 1) \right)^{-1}.$$

Furthermore, the identity $\Gamma(x + 1) = x\Gamma(x)$ implies

$$\begin{aligned}
\Gamma(\alpha_i + k_i + \mathbb{1}_{\{i \in A\}}) &= (\alpha_i + k_i)^{\mathbb{1}_{\{i \in A\}}} \Gamma(\alpha_i + k_i), \\
\Gamma(\beta_i + m_i - k_i + \mathbb{1}_{\{i \in B\}}) &= (\beta_i + m_i - k_i)^{\mathbb{1}_{\{i \in B\}}} \Gamma(\beta_i + m_i - k_i),
\end{aligned}$$

and thus both sides of (52) share a common factor of

$$\prod_{i=1}^{n} \Gamma(\alpha_i + k_i)\Gamma(\beta_i + m_i - k_i).$$

After cancelling out the common factors, (52) becomes equivalent to

$$\prod_{i\in A}(\alpha_i + k_i)\prod_{i\in B}(\beta_i + m_i - k_i) \quad > \quad \prod_{i\in B}(\alpha_i + k_i)\prod_{i\in A}(\beta_i + m_i - k_i),$$

which further simplifies to

$$\prod_{i\in A}\frac{\alpha_i + k_i}{\beta_i + m_i - k_i} > \prod_{i\in B}\frac{\alpha_i + k_i}{\beta_i + m_i - k_i}.$$

Hence, the choice (47) of $\hat{w}_i^{\mathrm{Ba}}$ guarantees that the decision rule in (46) is indeed optimal. ■

Remark 18 *Unfortunately, although*

$$\mathbb{P}(\hat{f}^{\mathrm{Ba}}(\mathbf{X}, \mathbf{K}) \neq Y) = \mathbb{P}(\hat{\mathbf{w}}^{\mathrm{Ba}} \cdot \boldsymbol{\eta} \leq 0)$$

is a deterministic function of $\{\alpha_i, \beta_i, m_i\}$, we are unable to compute it at this point, or even give a non-trivial bound. The main source of difficulty is the coupling between $\hat{\mathbf{w}}^{\mathrm{Ba}}$ and $\boldsymbol{\eta}$.

Open problem. Give a non-trivial estimate for $\mathbb{P}(\hat{f}^{\mathrm{Ba}}(\mathbf{X}, \mathbf{K}) \neq Y)$.

6. Experiments

It is most instructive to take the committee size n to be small when comparing the different voting rules. Indeed, for a large committee of "marginally competent" experts with $p_i = \frac{1}{2} + \gamma$ for some $\gamma > 0$, even the simple majority rule $f^{\mathrm{MAJ}}(\mathbf{x}) = \mathrm{sign}(\sum_{i=1}^{n} x_i)$ has a probability of error decaying as $\exp(-4n\gamma^2)$, as can be easily seen from Hoeffding's bounds. The more sophisticated voting rules discussed in this paper perform even better in this setting; see Helmbold and Long (2012) for an in-depth study of the utility gained from weak experts. Hence, small committees provide the natural test-bed for gauging a voting rule's ability to exploit highly competent experts. In our experiments, we set $n = 5$ and the sample sizes m_i were identical for all experts. The results were averaged over 10^5 trials. Two of our experiments are described below.

Low vs. high confidence. The goal of this experiment was to contrast the extremal behavior of $\hat{f}^{\mathrm{LC}}$ vs. $\hat{f}^{\mathrm{HC}}$. To this end, we numerically optimized the $\mathbf{p} \in [0,1]^n$ so as to maximize the *absolute gap*

$$\Delta_n(\mathbf{p}) := \mathbb{P}(f^{\mathrm{LC}}(\mathbf{X}) \neq Y) - \mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) \neq Y),$$

where $f^{\mathrm{LC}}(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{n}(p_i - \frac{1}{2})x_i\right)$. We were surprised to discover that, though the ratio $\mathbb{P}(f^{\mathrm{LC}}(\mathbf{X}) \neq Y)/\mathbb{P}(f^{\mathrm{OPT}}(\mathbf{X}) \neq Y)$ can be made arbitrarily large by setting $p_1 \approx 1$ and the remaining $p_i < 1 - \varepsilon$, the absolute gap appears to be rather small: we conjecture (with some heuristic justification[8]) that $\sup_{n\geq 1}\sup_{\mathbf{p}\in[0,1]^n}\Delta_n(\mathbf{p}) = 1/16$. For $\hat{f}^{\mathrm{Ba}}$, we used $\alpha_i = \beta_i = 1$ for all i. The results are reported in Figure 1.

8. The intuition is that we want one of the experts to be perfect (i.e., $p = 1$) and two others to be "moderately strong," whereby under the low confidence rule, the two can collude to overwhelm the perfect

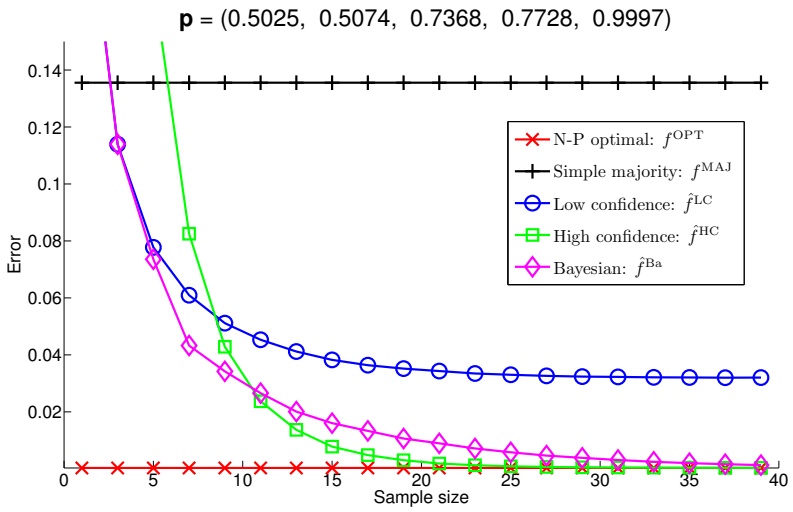

Figure 1: For very small sample sizes, $\hat{f}^{\mathrm{LC}}$ outperforms $\hat{f}^{\mathrm{HC}}$ but is outperformed by $\hat{f}^{\mathrm{Ba}}$. Starting from sample size ≈ 13, $\hat{f}^{\mathrm{HC}}$ dominates the other empirical rules. The empirical rules are (essentially) sandwiched between f^{OPT} and f^{MAJ}.

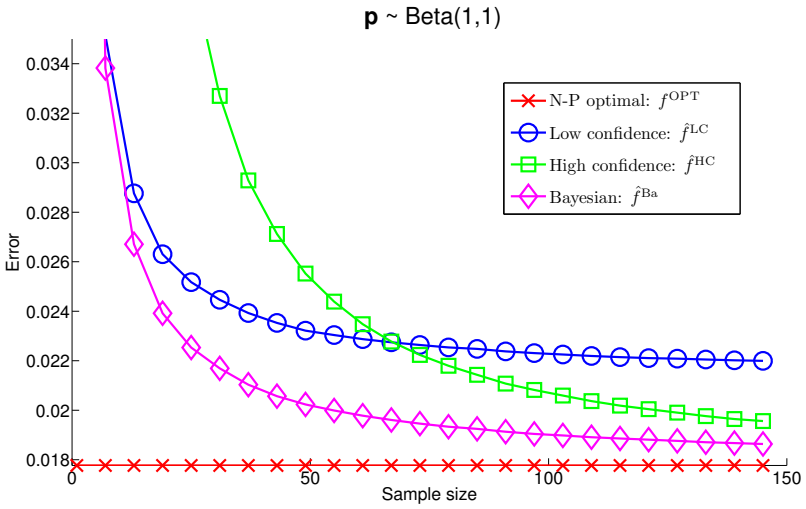

Figure 2: Unsurprisingly, $\hat{f}^{\mathrm{Ba}}$ uniformly outperforms the other two empirical rules. We found it somewhat surprising that $\hat{f}^{\mathrm{HC}}$ required so many samples (about 60 on average) to overtake $\hat{f}^{\mathrm{LC}}$. The simple majority rule f^{MAJ} (off the chart) performed at an average accuracy of 50%, as expected.

expert, but neither of them alone can. For $n = 3$, the choice $\mathbf{p} = (1, 3/4 + \varepsilon, 3/4 + \varepsilon)$ asymptotically achieves the gap $\Delta_3(\mathbf{p}) = 1/16$.

Bayesian setting. In each trial, a vector of expert competences $\mathbf{p} \in [0,1]^n$ was drawn independently componentwise, with $p_i \sim \mathrm{Beta}(1,1)$. These values (i.e., $\alpha_i = \beta_i \equiv 1$) were used for $\hat{f}^{\mathrm{Ba}}$. The results are reported in Figure 2.

7. Discussion

The classic and seemingly well-understood problem of the consistency of weighted majority votes continues to reveal untapped depth and suggest challenging unresolved questions. We hope that the results and open problems presented here will stimulate future research.

Acknowledgements

We thank Tony Jebara, Phil Long, Elchanan Mossel, and Boaz Nadler for enlightening discussions and for providing useful references. This paper greatly benefited from a careful reading by two diligent referees, who corrected inaccuracies and even supplied some new results. A special thanks to Lawrence Saul for writing up the new proof of the Kearns-Saul inequality and allowing us to print it here.

Appendix A. Bibliographical Notes on the Kearns-Saul Inequality

Given the recent interest surrounding the Kearns-Saul inequality (9), we find it instructive to provide some historical notes on this and related results. Most of the material in this section is taken from Saul (2014), to whom we are indebted for writing the note and for his kind permission to include it in this paper.

Lemma 19 *Let $f(x) = \log\cosh(\frac{1}{2}\sqrt{x})$. Then $f(x)$ is concave on $x \geq 0$.*

Proof The second derivative is given by

$$f''(x) \;=\; \frac{\mathrm{sech}^2(\frac{1}{2}\sqrt{x})}{16x^{3/2}}\left[\sqrt{x} - \sinh(\sqrt{x})\right].$$

For $x > 0$, the first of these factors is positive, and the second is negative. To show the latter, recall the Taylor series expansion

$$\sinh(t) = t + \frac{t^3}{3!} + \frac{t^5}{5!} + \frac{t^7}{7!} + \ldots,$$

from which we observe that $\sqrt{x} \leq \sinh(\sqrt{x})$. It also follows from the Taylor series that $f''(0) = -\frac{1}{96}$. It follows that f'' is negative on the positive half-line, and hence f is concave on this domain. ■

Corollary 20 *For $x, x_0 > 0$, we have*

$$\log\cosh(\tfrac{1}{2}\sqrt{x}) \;\leq\; \log\cosh(\tfrac{1}{2}\sqrt{x_0}) + \left[\frac{\tanh(\frac{1}{2}\sqrt{x_0})}{4\sqrt{x_0}}\right](x - x_0). \tag{53}$$

Proof A concave function $f(x)$ is upper-bounded by its first-order Taylor approximation: $f(x) \leq f(x_0) + f'(x_0)(x - x_0)$. The claim follows from Lemma 19. ■

The results in Lemma 19 and Corollary 20 were first stated by Jaakkola and Jordan (1997); see Jebara (2011); Jebara and Choromanska (2012) for extensions, including a multivariate version. As pointed out by a referee, Theorem 1 in Hoeffding (1963) contains some bounds that bear a resemblance to the Kearns-Saul inequality. However, we were unable to derive the latter from the former — which, in particular, requires all of the summands to be bounded between 0 and 1.

Suppose that in Equation (53), we make the substitutions

$$\sqrt{x} = \left| t + \log \frac{p}{1-p} \right|, \tag{54}$$

$$\sqrt{x_0} = \left| \log \frac{p}{1-p} \right|, \tag{55}$$

where $t \in \mathbb{R}$ and $p \in (0,1)$. Then we obtain a particular form of the bound that will be especially useful in what follows.

Corollary 21 *For all $t \in \mathbb{R}$ and $p \in (0,1)$,*

$$\log \cosh \left(\tfrac{1}{2} \left[t + \log \frac{p}{1-p} \right] \right) \leq -\log \left[2\sqrt{p(1-p)} \right] + (p - \tfrac{1}{2})t + \left(\frac{2p-1}{4 \log \frac{p}{1-p}} \right) t^2.$$

Proof Make the substitutions suggested in (54, 55) and apply Corollary 20. The result follows from tedious but elementary algebra. ■

The above result yields perhaps the most natural and direct proof of the Kearns-Saul inequality to date:

Theorem 22 *For all $t \in \mathbb{R}$ and $p \in (0,1)$,*

$$\log \left[(1-p)e^{-pt} + pe^{(1-p)t} \right] \leq \left(\frac{2p-1}{4 \log \frac{p}{1-p}} \right) t^2.$$

Proof Rewrite the left-hand side by symmetrizing the argument inside the logarithm,

$$\log \left[(1-p)e^{-pt} + pe^{(1-p)t} \right] = \log \cosh \left(\tfrac{1}{2} \left[t + \log \frac{p}{1-p} \right] \right) - (p - \tfrac{1}{2})t + \log \left[2\sqrt{p(1-p)} \right],$$

and invoke Corollary 21. ■

The inequality in Theorem 22 was first stated by Kearns and Saul (1998) and first rigorously proved by Berend and Kontorovich (2013b). Shortly thereafter, Raginsky (2012) provided a very elegant proof based on transportation and information-theoretic techniques, which currently appears as Theorem 37 in Raginsky and Sason (2013). A third proof, found by Schlemm (2014), fleshes out the original strategy suggested by Kearns and Saul (1998). The fourth proof, given here, is due to Saul (2014).

References

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory (ALT)*, 2007.

Eyal Baharad, Jacob Goldberger, Moshe Koppel, and Shmuel Nitzan. Distilling the wisdom of crowds: weighted aggregation of decisions on multiple issues. *Autonomous Agents and Multi-Agent Systems*, 22(1):31–42, 2011.

Eyal Baharad, Jacob Goldberger, Moshe Koppel, and Shmuel Nitzan. Beyond Condorcet: Optimal aggregation rules using voting records. *Theory and Decision*, 72(1):113–130, 2012.

Daniel Berend and Aryeh Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013a.

Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electron. Commun. Probab.*, 18:no. 3, 1–7, 2013b.

Daniel Berend and Aryeh Kontorovich. Consistency of weighted majority votes. In *Neural Information Processing Systems (NIPS)*, 2014.

Daniel Berend and Jacob Paroush. When is Condorcet's jury theorem valid? *Soc. Choice Welfare*, 15(4):481–488, 1998.

Daniel Berend and Luba Sapir. Monotonicity in Condorcet's jury theorem with dependent voters. *Social Choice and Welfare*, 28(3):507–528, 2007.

Daniel Berend, Peter Harremoës, and Aryeh Kontorovich. Minimum KL-divergence on complements of L_1 balls. *IEEE Transactions on Information Theory*, 60(6):3172–3177, 2014.

Philip J. Boland, Frank Proschan, and Y. L. Tong. Modelling dependence in simple and indirect majority systems. *J. Appl. Probab.*, 26(1):81–88, 1989. ISSN 0021-9002.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, NJ, second edition, 2006.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979.

J.A.N. de Caritat marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. AMS Chelsea Publishing Series. Chelsea Publishing Company, 1785.

Frank den Hollander. *Large deviations*, volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2000.

Elad Eban, Elad Mezuman, and Amir Globerson. Discrete chebyshev classifiers. In *International Conference on Machine Learning (ICML) (2)*, 2014.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. The value of observation for monitoring dynamic systems. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2007.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Chao Gao and Dengyong Zhou. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels (arxiv:1310.5764). 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2009.

David P. Helmbold and Philip M. Long. On the necessity of irrelevant variables. *Journal of Machine Learning Research*, 13:2145–2170, 2012.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

Tommi S. Jaakkola and Michael I. Jordan. A variational approach to Bayesian logistic regression models and their extensions. In *Artificial Intelligence and Statistics, AISTATS*, 1997.

Tony Jebara. Multitask sparsity via maximum entropy discrimination. *Journal of Machine Learning Research*, 12:75–110, 2011.

Tony Jebara and Anna Choromanska. Majorization for CRFs and latent likelihoods. In *Neural Information Processing Systems (NIPS)*, 2012.

Michael J. Kearns and Lawrence K. Saul. Large deviation methods for approximate probabilistic inference. In *Uncertainty in Artificial Intelligence (UAI)*, 1998.

Aryeh Kontorovich. Obtaining measure concentration from Markov contraction. *Markov Processes and Related Fields*, 4:613–638, 2012.

Aryeh (Leonid) Kontorovich. *Measure Concentration of Strongly Mixing Processes with Applications*. PhD thesis, Carnegie Mellon University, 2007.

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Neural Information Processing Systems (NIPS)*, 2006.

François Laviolette and Mario Marchand. PAC-Bayes risk bounds for stochastic averages and majority votes of sample-compressed classifiers. *Journal of Machine Learning Research*, 8:1461–1487, 2007.

Hongwei Li, Bin Yu, and Dengyong Zhou. Error rate bounds in crowdsourcing models. *CoRR*, abs/1307.2674, 2013.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. In *Foundations of Computer Science (FOCS)*, 1989.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, 1994.

Yishay Mansour, Aviad Rubinstein, and Moshe Tennenholtz. Robust aggregation of experts signals. 2013.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Conference on Learning Theory (COLT)*, 2009.

David A. McAllester and Luis E. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895–911, 2003.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *International Conference on Machine Learning (ICML)*, 2008.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337, 1933.

Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–297, 1982.

Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, 2014.

Maxim Raginsky. Derivation of the Kearns-Saul inequality by optimal transportation (private communication), 2012.

Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–247, 2013.

Jean-Francis Roy, François Laviolette, and Mario Marchand. From PAC-Bayes bounds to quadratic programs for majority votes. In *International Conference on Machine Learning (ICML)*, 2011.

Lawrence K. Saul. Yet another proof of an obscure inequality (private communication), 2014.

Robert E. Schapire and Yoav Freund. *Boosting. Foundations and algorithms.* Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2012.

Eckhard Schlemm. The Kearns–Saul inequality for Bernoulli and Poisson-binomial distributions. *Journal of Theoretical Probability*, pages 1–15, 2014.