EXACT UPPER AND LOWER BOUNDS ON THE MISCLASSIFICATION PROBABILITY

IOSIF PINELIS

ABSTRACT. Exact lower and upper bounds on the best possible misclassification probability for a finite number of classes are obtained in terms of the total variation norms of the differences between the sub-distributions over the classes. These bounds are compared with the exact bounds in terms of the conditional entropy obtained by Feder and Merhav.

CONTENTS

1. Introduction, summary and discussion

Let X and Y be random variables (r.v.'s) defined on the same probability space $(\Omega, \mathscr{F}, \mathsf{P})$, X with values in a set S (endowed with a sigma-algebra Σ) and Y with values in the set $[k] := \{1, \ldots, k\}$, where k is a natural number; to avoid trivialities, assume $k \geqslant 2$.

The sets Ω and $[k]$ may be regarded, respectively, as the population of objects of interest and the set of all possible classification labels for those objects. For each "object" $\omega \in \Omega$, the corresponding values $X(\omega) \in S$ and $Y(\omega) \in [k]$ of the r.v.'s X and Y may be interpreted as the (correct) description of ω and the (correct) classification label for ω, respectively.

Alternatively, $Y(\omega)$ may be interpreted as the signal entered at the input side of a device – with its possibly corrupted, output version $X(\omega)$.

The problem is to find a good or, better, optimal way to reconstruct, for each $\omega \in \Omega$, the correct label (or input signal) $Y(\omega)$ based on the description (or, respectively, the output signal) $X(\omega)$. To solve this problem, one uses a measurable function $f \colon S \to [k]$, referred to as a classification rule or, briefly, a classifier, which assigns a label (an input signal) $f(x) \in [k]$ to each possible description (or, respectively, to each possible output signal) $x \in S$. Then

$$p_f := \mathsf{P}(f(X) \neq Y)$$

is the misclassification probability for the classifier f.

2010 *Mathematics Subject Classification.* Primary: 60E15, 62H30, 91B06. Secondary: 26D15, 26D20, 62C10, 68T10.

Key words and phrases. Classification problem, misclassification probability, total variation norm, Bayes estimators, maximum likelihood estimators.

For each $y \in [k]$, let μ_y be the sub-probability measure on Σ defined by the condition

(1) $$\mu_y(B) := \mathsf{P}(Y = y, X \in B)$$

for $B \in \Sigma$, so that

(2) $$\mu := \mu_1 + \cdots + \mu_k$$

is the probability measure that is the distribution of X in S, and let

$$\rho_y := \frac{\mathrm{d}\mu_y}{\mathrm{d}\mu},$$

the density of μ_y with respect to μ.

The value $y \in [k]$ may be considered a parameter, so that the problem may be be viewed as one of Bayesian estimation (of a discrete parameter, with values in the finite set $[k]$). If the r.v. X is discrete as well, then of course

$$\rho_y(x) = \mathsf{P}(Y = y | X = x)$$

for each $x \in S$ with $\mathsf{P}(X = x) \neq 0$. So, for each such x, the function $y \mapsto \rho_y(x)$ may be referred to as the probability mass function of the posterior distribution of the parameter corresponding to the observation x.

The following proposition is, essentially, a well-known fact of Bayesian estimation:

Proposition 1. *For each* $x \in S$, *let* $f_*(x) := \min \mathrm{argmax}_y \rho_y(x)$, *where* $\mathrm{argmax}_y \rho_y(x) := \{y \in [k] : \rho_y(x) = \max_{z \in [k]} \rho_z(x)\}$; *thus,* $f_*(x)$ *is the smallest maximizer of* $\rho_y(x)$ *in* $y \in [k]$. *Then the function* f_* *is a classifier, and*

$$p_* := p_{f_*} = 1 - \int_S \max_{y=1}^{k} \rho_y(x)\, \mu(\mathrm{d}x) \leqslant p_f$$

for any classifier f, *so that* p_* *is the smallest possible misclassification probability.*

The proofs of all statements that may need a proof are deferred to Section 2.

Let

(3) $$\Delta := \sum_{1 \leqslant y < z \leqslant k} \|\mu_y - \mu_z\| = \sum_{1 \leqslant y < z \leqslant k} |\rho_y - \rho_z| \mathrm{d}\mu,$$

where $\|\cdot\|$ is the total variation norm.

In the "population" model, the measure μ_y conveys two kinds of information: (i) the relative size $\frac{\|\mu_y\|}{\|\mu\|} = \|\mu_y\|$ (of the set of all individual descriptions) of the yth subpopulation (of the entire population Ω) consisting of the objects that carry the label y and (ii) the (conditional) probability distribution $\frac{\mu_y}{\|\mu_y\|}$ of the object descriptions in this yth subpopulation, assuming the size $\|\mu_y\|$ of the yth subpopulation is nonzero. Everywhere here, y and z are in the set $[k]$. Thus, Δ is a summary characteristic of the pairwise differences between the k subpopulations, which takes into account both of the two just mentioned kinds of information.

In the input-output model, the $\|\mu_y\|$'s are interpreted as the prior probabilities of the possible input signals $y \in [k]$ – whereas, for each $y \in [k]$, the (conditional) probability distribution $\frac{\mu_y}{\|\mu_y\|}$ is the distribution of the output signal corresponding to the given input y. Thus, here Δ is a summary characteristic of the pairwise differences between the k sets of possible outputs corresponding to the k possible inputs.

Remark 2. By (3),

$$0 \leqslant \Delta \leqslant \sum_{1 \leqslant y < z \leqslant k} (\|\mu_y\| + \|\mu_z\|) = (k-1) \sum_{1}^{k} \|\mu_y\| = k - 1.$$

Moreover, the extreme values 0 and $k-1$ of Δ are attained, respectively, when the measures μ_y are the same for all $y \in [k]$ and when these measures are pairwise mutually singular.

The main result of this paper provides the following upper and lower bounds on the smallest possible misclassification probability p_* in terms of Δ:

Theorem 3. *One has*

(4) $$L(\Delta) \leqslant p_* \leqslant U(\Delta) \leqslant U_{\mathsf{simpl}}(\Delta),$$

where

$$L(\Delta) := L_k(\Delta) := 1 - \frac{1 + \Delta}{k},$$

(5) $$U(\Delta) := U_k(\Delta) := 1 - \frac{k + 1 + \Delta - 2\lceil \Delta \rceil}{(k - \lceil \Delta \rceil)(k + 1 - \lceil \Delta \rceil)},$$

$$U_{\mathsf{simpl}}(\Delta) := U_{k;\mathsf{simpl}}(\Delta) := 1 - \frac{1}{k - \Delta},$$

and $\lceil \cdot \rceil$ is the ceiling function, so that $\lceil \Delta \rceil$ is the smallest integer that is no less than Δ.

Theorem 3 is complemented by

Proposition 4. *For each possible value of Δ in the interval $[0, k-1]$, the lower and upper bounds $L(\Delta)$ and $U(\Delta)$ on p_* are exact: For each $\Delta \in [0, k-1]$, there are r.v.'s X and Y as described in the beginning of this paper for which one has the equality $p_* = L(\Delta)$; similarly, with $U(\Delta)$ in place of $L(\Delta)$. More specifically, the first (respectively, second) inequality in (4) turns into the equality if and only if there is a set $S_0 \in \Sigma$ such that $\mu(S_0) = 0$ and for each $x \in S \setminus S_0$ the values $\rho_1(x), \ldots, \rho_k(x)$ constitute a permutation of numbers $a_1, \ldots, a_k$ as in (17) (respectively, in (18)) with $d = \Delta$. The simpler/simplified upper bound $U_{\mathsf{simpl}}(\Delta)$ is exact only for the integral values of Δ.*

Remark 5. In view of Remark 2, the functions L, U, and U_{simpl}, introduced in Theorem 3, are well defined on the interval $[0, k-1]$. Moreover, $U(\Delta)$ is the linear interpolation of $U_{\mathsf{simpl}}(\Delta)$ over the possible integral values $0, \ldots, k-1$ of Δ. Thus, each of the functions L, U, and U_{simpl} is concave and strictly decreasing (from $1 - \frac{1}{k}$ to 0) on the interval $[0; k-1]$; moreover, the function L is obviously affine. We see that, the greater is the characteristic Δ of the pairwise differences between the k subpopulations, the smaller are the lower and upper bounds $L(\Delta)$, $U(\Delta)$, and $U_{\mathsf{simpl}}(\Delta)$ on the misclassification probability p_*. Of course, this quite corresponds to what should be expected of good bounds on p_*. It also follows that one always has

(6) $$0 \leqslant p_* \leqslant 1 - \frac{1}{k},$$

and the extreme values 0 and $1 - \frac{1}{k}$ of the misclassification probability p_* are attained when, respectively, $\Delta = k-1$ and $\Delta = 0$. The bounds L, U, and U_{simpl} are illustrated in Figure 1.

FIGURE 1. Graphs of the bounds L (green), U (green), and U_{simpl} (dark green) for $k = 5$.

Feder and Merhav [3] obtained the following exact upper and lower bounds of the optimal misclassification probability in terms of the conditional entropy H:

$$L_{\mathsf{FM}}(H) \leqslant p_* \leqslant U_{\mathsf{FM}}(H),$$

where

$$(7) \qquad H := H(Y|X) := -\mathsf{E}\sum_{y=1}^{k} \rho_y(X)\ln\rho_y(X) = -\int_S \mu(\mathrm{d}x)\sum_{y=1}^{k} \rho_y(x)\ln\rho_y(x),$$

(8)
$$L_{\mathsf{FM}}(H) := \Phi^{-1}(H), \quad \Phi(p) := p\ln(k-1)+h_2(p), \quad h_2(p) := -p\ln p-(1-p)\ln(1-p)$$

for $p \in (0, 1)$, $h_2(0) := 0$, $h_2(1) := 0$,

$$(9) \qquad U_{\mathsf{FM}}(H) := \frac{e(H)-1}{e(H)} + \frac{1}{e(H)(e(H)+1)}\frac{H-\ln e(H)}{\ln(1+1/e(H))},$$

and

$$(10) \qquad e(H) := \lceil e^H \rceil - 1.$$

Note that $\Phi(p)$ strictly and continuously increases from 0 to $\ln k$ as p increases from 0 to $1 - \frac{1}{k}$. Therefore and because all the values of the conditional entropy H lie between 0 and $\ln k$, the expression $\Phi^{-1}(H)$ is well defined, and its values lie between 0 and $1 - \frac{1}{k}$ – which is in accordance with (6).

Throughout this paper, we use only natural, base-e logarithms. In [3], the bounds are stated in terms of binary, base-2 logarithms. To rewrite $L_{\mathsf{FM}}(H)$ and $U_{\mathsf{FM}}(H)$ in terms of binary logarithms, replace all the instances of $\ln = \log_e$ in (7)–(9) by $\log_2$ and, respectively, replace e^H in (10) by 2^H. An advantage of using natural logarithms is that then the expressions for the corresponding derivatives, used in our proofs, are a bit simpler; also, $\ln$ is a bit shorter in writing than $\log_2$ or even $\log$.

Note also that, in the notation in [3], the roles of X and Y are reversed: there, X denotes the input and Y the output. Our notation in this paper is in accordance with the standard convention in machine learning; cf. e.g. [5, 4].

Let us compare, in detail, our "Δ-bounds" $L(\Delta)$, $U(\Delta)$, and $U_{\mathsf{simpl}}(\Delta)$ with the "H-bounds" $L_{\mathsf{FM}}(H)$ and $U_{\mathsf{FM}}(H)$. We shall be making the comparisons only in

the "pure" settings, when the set $\{\rho_1(x), \ldots, \rho_k(x)\}$ is the same for all $x \in S$, that is, when for each $x \in S$ the k-tuple $(\rho_1(x), \ldots, \rho_k(x))$ is a permutation of one and the same k-tuple $(a_1, \ldots, a_k)$ (of nonnegative real numbers $a_1, \ldots, a_k$ such that $a_1 + \cdots + a_k = 1$). A reason for doing so is that one may expect the comparisons to be of greater contrast in the "pure" settings than in "mixed", non-"pure" ones. Thus, focusing on "pure" settings will likely allow us to see the differences between the "Δ-bounds" and the "H-bounds" more clearly, while taking less time and effort.

We shall see that, even though the "H-bounds" $L_{\mathsf{FM}}(H)$ and $U_{\mathsf{FM}}(H)$ and the "Δ-bounds" $L(\Delta)$ and $U(\Delta)$ are exact in terms of H and Δ, respectively, they have rather different properties.

Remark 6. Typically, the lower H-bound $L_{\mathsf{FM}}(H)$ on p_* appears to be better (that is, larger) than the lower Δ-bound $L(\Delta)$, whereas the upper H-bound $U_{\mathsf{FM}}(H)$ on p_* appears to be worse (that is, larger) than the upper Δ-bound $U(\Delta)$ and even its simplified but less accurate version $U_{\mathsf{simpl}}(\Delta)$.

However, in some rather exceptional cases these relations are reversed.

In particular, if the best possible misclassification probability p_* is large enough, then the lower Δ-bound $L(\Delta)$ may be better than the lower H-bound $L_{\mathsf{FM}}(H)$, for each $k \geqslant 3$.

On the other hand, if k is large enough and p_* is small enough, then the upper Δ-bound $U(\Delta)$ may be worse than the upper H-bound $U_{\mathsf{FM}}(H)$. However, I have not been able to find cases with $U(\Delta)$ (or even $U_{\mathsf{simpl}}(\Delta)$) worse than $U_{\mathsf{FM}}(H)$ when there are at most $k = 9$ classes.

More specifically, we have the following propositions. (As usual, $\mathrm{I}\{\cdot\}$ will denote the indicator function.)

Proposition 7. *Suppose that $k \geqslant 3$ and for each $x \in S$ the vector $(\rho_1(x), \ldots, \rho_k(x))$ is a permutation of the vector $(a_1, \ldots, a_k)$, where*

$$a_i = \frac{1}{\ell} \, \mathrm{I}\{1 \leqslant i \leqslant \ell\}$$

for some natural $\ell \geqslant 2$ in the set $\{k-3, k-2, k-1\}$ and for all $i = 1, \ldots, k$; one may also allow $\ell = k - 4$ if $k \in \{6, 7, 8, 9\}$. Then $L(\Delta) > L_{\mathsf{FM}}(H)$.

Proposition 8. *Fix any $\nu \in (1, \infty)$. Suppose that for each $x \in S$ the vector $(\rho_1(x), \ldots, \rho_k(x))$ is a permutation of the vector $(a_1, \ldots, a_k)$, where $k > \nu$ and*

$$a_i = \left(1 - \frac{\nu - 1}{k}\right) \mathrm{I}\{i = 1\} + \frac{\nu - 1}{k(k-1)} \, \mathrm{I}\{2 \leqslant i \leqslant k\}$$

for all $i = 1, \ldots, k$. Then $U(\Delta) > U_{\mathsf{FM}}(H)$ for all large enough k (depending on the value of ν).

Note that in Proposition 7 the best possible misclassification probability $p_* = 1 - \frac{1}{\ell}$ is large, especially when ℓ is large (and hence so is k). In contrast, in Proposition 8 $p_* = \frac{\nu - 1}{k}$ is small for the large values of k, assumed in that proposition. Either of these two kinds of situations, especially the second one, may be considered somewhat atypical: it usually should be difficult to make the misclassification probability p_* small when the number k of possible classes is large; on the other hand, when k is not very large, one may hope that the best possible misclassification probability is small enough.

Concerning the case of two classes, we have

Proposition 9. *Suppose that $k = 2$. Then $U(\Delta) = L(\Delta) = L_{\mathsf{FM}}(H) = p_*$ for all pairs of r.v.'s (X, Y). So, one can say that the bounds $U(\Delta)$, $L(\Delta)$, and $L_{\mathsf{FM}}(H)$ always perfectly estimate the best possible misclassification probability p_* – if $k = 2$.*

On the other hand, here $U_{\mathsf{FM}}(H) > U_{\mathsf{simpl}}(\Delta) > p_*$ unless there is a set $S_0 \in \Sigma$ such that $\mu(S_0) = 0$ and for each $x \in S \setminus S_0$ either $\rho_1(x) = \rho_2(x) = 1/2$ or $\{\rho_1(x), \rho_2(x)\} = \{0, 1\}$ – that is, the values $\rho_1(x)$ and $\rho_2(x)$ constitute a permutation of the numbers 0 and 1. Thus, in the case $k = 2$, with the mentioned trivial exceptions, even the simplified upper Δ-bound $U_{\mathsf{simpl}}(\Delta)$ on p_* is strictly better than the upper H-bound $U_{\mathsf{FM}}(H)$, but still $U_{\mathsf{simpl}}(\Delta)$ is not a perfect estimate of p_*.

An important case is that of three classes, so that $k = 3$. Here, in the "pure" setting, for each $x \in S$ the triple $(\rho_1(x), \rho_2(x), \rho_2(x))$ is a permutation of the triple $(1 - p, p - \varepsilon, \varepsilon)$, where $p := p_* \in [0, 1 - 1/3]$ and $1 - p \geqslant p - \varepsilon \geqslant \varepsilon \geqslant 0$ or, equivalently, $p \in [0, 2/3]$ and $(2p - 1)_+ \leqslant \varepsilon \leqslant p/2$, where $u_+ := \max(0, u)$. Each of the 6 pictures in Figure 2 presents the graphs of the decimal logarithms of the bounds $L(\Delta)$, $U(\Delta)$, $U_{\mathsf{simpl}}(\Delta)$, $L_{\mathsf{FM}}(H)$, and $U_{\mathsf{FM}}(H)$ as functions of $\varepsilon \in [(2p - 1)_+, p/2]$ with the misclassification probability $p = p_*$ taking a fixed value in the set $\{0.01, 0.1, 0.3, 0.5, 0.6, 0.64\}$. We see that in all these cases the upper Δ-bound $U(\Delta)$ and even its simplified (but worse) version $U_{\mathsf{simpl}}(\Delta)$ are better than the upper H-bound $U_{\mathsf{FM}}(H)$, over the entire range of values of ε. For small values of the best possible misclassification probability p_*, the lower H-bound $L_{\mathsf{FM}}(H)$ is significantly better than $L(\Delta)$ over all values of ε; however, this comparison is reversed if p_* is large enough but ε is small enough (especially in the case $p_* = 0.5$).

An interesting series of cases is given by what may be called the binomial model (with a parameter $q \in (0, 1)$), in which $k = 2^m$ for a natural m, and for each $x \in S$ the vector $(\rho_1(x), \ldots, \rho_k(x))$ is a permutation of a vector $(a_1, \ldots, a_k)$, where each a_i is of the form $(1 - q)^j q^{m-j}$ for some $j \in \{0, \ldots, m\}$, and the multiplicity of the form $(1 - q)^j q^{m-j}$ among the a_i's is $\binom{m}{j}$ for each $j \in \{0, \ldots, m\}$. Clearly then, all the a_i's are nonnegative, and $a_1 + \cdots + a_k = \sum_{j=0}^{m} \binom{m}{j}(1 - q)^j q^{m-j} = 1$. In particular, for $m = 1$ we have $k = 2$, and then we may take $(a_1, a_2) = (1 - q, q)$. For $m = 2$ we have $k = 4$, and then we may take

$$(a_1, a_2, a_3, a_4) = \big((1 - q)^2, \, (1 - q)q, \, (1 - q)q, \, q^2\big).$$

Choosing, in the latter case, $S = \{1, 2, 3, 4\}$ and $q = Q(\sqrt{2E_b/N_0})$, where $Q(x) := \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ is the tail probability for the standard normal distribution, E_b is the energy per bit, and $N_0/2$ is the noise power spectral density (PSD), we see that the resulting particular case of the binomial model covers the so-called quadrature phase-shift keying (QPSK) digital communication scheme over an additive white Gaussian noise (AWGN) channel (cf. e.g. [2, page 313]), which in fact provided the motivation for the general binomial model.

Another interesting series of cases is given by what may be called the exponential model (with a parameter $q \in (0, 1)$), in which for each $x \in S$ the vector $(\rho_1(x), \ldots, \rho_k(x))$ is a permutation of a vector $(a_1, \ldots, a_k)$, where $a_i := (1 - q)^{i-1} q^{k-i}/c_q$ and $c_q := c_{k,q} := \sum_{i=1}^{k} (1 - q)^{i-1} q^{k-i}$, so that all the a_i's are nonnegative and $a_1 + \cdots + a_k = 1$. Informally, the exponential model can be obtained from the binomial one by removing the multiplicities.

Graphical comparisons of the "Δ-bounds" with the "H-bounds" (as functions of the parameter q) for the cases $k = 2, 4, 8$ of the binomial and exponential models

are presented in Figure 3. Note here that, by symmetry, it is enough to consider $q \in (0, 1/2]$. Obviously, for $k = 2$ the binomial and exponential models are the same, and in this case they are the same as the essentially unique general "pure" model for $k = 2$, fully considered in Proposition 9. Accordingly, the pictures in the first row in Figure 3 are identical to each other, and the graphs of the bounds $U(\Delta)$, $L(\Delta)$, and $L_{\mathsf{FM}}(H)$ are the same as that of p_*. The cases $k = 4, 8$ in Figure 3 illustrate the first sentence in Remark 6. It appears that the comparisons in the exponential model are somewhat more favorable to the Δ-bounds than they are in the binomial model.

FIGURE 2. Graphs of $\log_{10} L(\Delta)$ (green), $\log_{10} U(\Delta)$ (green), $\log_{10} U_{\mathsf{simpl}}(\Delta)$ (dark green), $\log_{10} L_{\mathsf{FM}}(H)$ (blue), $\log_{10} U_{\mathsf{FM}}(H)$ (blue), and $\log_{10} p_*$ (dashed) for $k = 3$ and $p_* \in \{0.01, 0.1, 0.3, 0.5, 0.6, 0.64\}$.

FIGURE 3. Graphs of $\log_{10} L(\Delta)$ (green), $\log_{10} U(\Delta)$ (green), $\log_{10} U_{\mathsf{simpl}}(\Delta)$ (dark green), $\log_{10} L_{\mathsf{FM}}(H)$ (blue), $\log_{10} U_{\mathsf{FM}}(H)$ (blue), and $\log_{10} p_*$ (dashed) for $k = 2, 4, 8$. Left column: Binomial model. Right column: Exponential model.

Upper and lower bounds on the best possible misclassification probability in terms of Renyi's conditional entropy were announced in [1], where the input and

output r.v.'s were denoted by M and W, respectively, and both M and W were assumed to take values in the same set $\{1, \ldots, k\}$. Renyi's conditional entropy used in [1] was defined by the formula

$$(11) \qquad H_\beta(W|M) := \sum_{m=1}^{k} \mathsf{P}(M = m) H_\beta(W|m),$$

where

$$(12) \qquad H_\beta(W|m) := \frac{1}{1 - \beta} \log \sum_{w=1}^{k} \mathsf{P}(W = w|M = m)^\beta, \quad \log := \log_2,$$

and $\beta \in (0,1) \cup (1,\infty)$, so that Shannon's conditional entropy

$$(13) \qquad H_1(W|M) := H(W|M) = - \sum_{m=1}^{k} \mathsf{P}(W = w) \log \mathsf{P}(W = w|M = m)$$

may be viewed as a limit case of Renyi's: $H_\beta(W|M) \to H(W|M)$ as $\beta \to 1$. Note that $H_\beta(W|M)$ is the conditional entropy of the output W given the input M. Thus, for some reason, the standard roles of the input and output r.v.'s were reversed in [1]; cf. e.g. the conditional entropy *of the input given the output* used in [3, formula (3)].

The mentioned upper and lower bounds announced in [1] were given by inequalities of the form

$$(14) \qquad \frac{H_\alpha(W|M) - H_S(e)}{N_1} \leqslant P(e) \leqslant \frac{H_\beta(W|M) - H_S(e)}{N_2},$$

where $\beta < 1 \leqslant \alpha$, N_1, N_2 are some positive expressions,

$$H_S(e) := -P(e) \log P(e) - (1 - P(e)) \log(1 - P(e)),$$

and $P(e)$ is the "the classification error probability", which doers not seem to be explicitly defined in [1]. A proof of these bounds was offered later in [2, Appendix], from which it appears (see [2, formula (A.1)]) that $P(e)$ is understood as $\mathsf{P}(W \neq M)$. However, the proof in [2] of the upper bound on $P(e)$ in (14) appears to be mistaken, and the upper bound itself may be negative and thus false in general.

Indeed, the second inequality in (14) appears to be obtained in [2] by multiplying the expressions in [2, formula (A.7)] by $p(m_k)$, then summing in k (in the notations there), and finally using the inequality $\sum_k p(m_k) H_S(e|m_k)[= H_S(e|M)] \geqslant H_S(e)$. However, in general the reverse inequality is true: $H_S(e|M) \leqslant H_S(e)$; cf. even the derivation of (A.6) from (A.5) in [2]. Also, the proof in [2] does not use the condition that $P(e)$ is *the smallest possible* classification error probability, and without such a condition no reasonable upper bound on $P(e)$ is possible.

More importantly, as mentioned above, the upper bound on $P(e)$ in (14) is false in general. For a very simple counterexample, suppose that $k = 2$, $\mathsf{P}(W = 1, M = 1) = \mathsf{P}(W = 1, M = 2) = 1/2$, and $\mathsf{P}(W = 2, M = 1) = \mathsf{P}(W = 2, M = 2) = 0$. Then $P(e) = 1/2$, $H_\beta(W|M) = 0$ for all β, and $H_S(e) = 1$, so that the presumed upper bound on $P(e)$ in (14) is negative and thus false.

Another two pairs of upper and lower bounds were announced in [1], in terms of the joint Renyi's joint entropy $H_\beta(W, M)$ of (W, M) and Renyi's mutual information $I_\beta(W, M)$ between W and M, rather than Renyi's conditional entropy, with no apparent proofs for these additional bounds. However, the same simple example

given above will quite similarly show that these additional upper bounds are false in general, too.

As stated in the abstract in [2], the mentioned bounds in [1] on $P(e)$ "were practically incomputable", because $P(e)$ itself appears in those bounds. Therefore, an effort was made in [2] to modify the bounds in [1] – by making the upper bounds greater and the lower bounds smaller – to make them computable. However, in view of what has been said, the modified upper bounds in [2] remain without a valid proof. As for the modified lower bounds, it is stated in [2, page 313] that, in the examples considered there, "the modified lower bounds are not depicted because they turn out to be negative".

For all these reasons, we shall not attempt to compare our bounds with ones in [1, 2].

2. Proofs

Proof of Proposition 1. Clearly, f_* is a map from S to $[k]$. Also, f_* is measurable, since $f_*^{-1}(\{y\}) = B_y \setminus \bigcup_{z=1}^{y-1} B_z \in \Sigma$ for each $y \in [k]$, where $B_y := \bigcap_{z=1}^{k} B_{y,z}$ and $B_{y,z} := \{x \in S \colon \rho_y(x) \geqslant \rho_z(x)\} \in \Sigma$. Thus, f_* is a classifier. Moreover, for any classifier f,

$$1 - p_f = \mathsf{P}(f(X) = Y) = \sum_{y=1}^{k} \mathsf{P}(Y = y, f(X) = y)$$

$$= \sum_{y=1}^{k} \int_S \mathrm{I}\{f(x) = y\} \, \mu_y(\mathrm{d}x)$$

$$= \sum_{y=1}^{k} \int_S \mathrm{I}\{f(x) = y\} \, \rho_y(x) \, \mu(\mathrm{d}x)$$

$$= \int_S \sum_{y=1}^{k} \mathrm{I}\{f(x) = y\} \, \rho_y(x) \, \mu(\mathrm{d}x)$$

$$\leqslant \int_S \max_{y=1}^{k} \rho_y(x) \, \mu(\mathrm{d}x) = 1 - p_{f_*}.$$

This completes the proof of Proposition 1. □

In view of Proposition 1 and (3), Theorem 3 and Proposition 4 follow immediately by the lemma below, with $\rho_i(x)$ in place of a_i.

Lemma 10. *Suppose that*

$$(15) \qquad a_1, \ldots, a_k \text{ are nonnegative real numbers such that } \sum_{1}^{k} a_i = 1.$$

Then

$$(16) \qquad L(\delta) \leqslant 1 - \max_{1}^{k} a_i \leqslant U(\delta) \leqslant U_{\mathsf{simpl}}(\delta),$$

where

$$\delta := \sum_{1 \leqslant i < j \leqslant k} |a_i - a_j|$$

and the functions L, U, and U_{simpl} are defined as in Theorem 3.

Under the stated conditions on the a_i's, one always has $0 \leqslant \delta \leqslant k - 1$; cf. Remark 2.

The bounds $L(\delta)$ and $U(\delta)$ on $1 - \max\limits_{1}^{k} a_i$ are exact for each possible value of δ:

(i) For each $d \in [0, k - 1]$, if

(17) $$a_1 = \frac{1 + d}{k} \quad and \quad a_2 = \cdots = a_k = \frac{1}{k} - \frac{d}{k(k - 1)},$$

then condition (15) holds, $\delta = d$, $\left[\max\limits_{1}^{k} a_i = a_1,\right]$ and the first inequality in (16) turns into the equality. If the a_i's satisfy condition (15) but do not constitute a permutation of the a_i's as in (17) with $d = \delta$, then the first inequality in (16) is strict.

(ii) For each $d \in [0, k - 1]$, if

(18) $$a_i = (1 - U(d)) \, \mathrm{I}\{i \leqslant k - \lceil d \rceil\} + \frac{\lceil d \rceil - d}{k + 1 - \lceil d \rceil} \, \mathrm{I}\{i = k + 1 - \lceil d \rceil\}$$

for all $i \in [k]$, then condition (15) holds, $\delta = d$, $\left[\max\limits_{1}^{k} a_i = a_1,\right]$ and the second inequality in (16) turns into the equality. If the a_i's satisfy condition (15) but do not constitute a permutation of the a_i's as in (18) with $d = \delta$, then the second inequality in (16) is strict.

Proof. It is quite easy to see that $U_{\mathsf{simpl}}(d)$ is concave in $d \in [0, k - 1]$. Moreover, as noted in Remark 5, $U(d)$ is the linear interpolation of $U_{\mathsf{simpl}}(d)$ over $d = 0, \ldots, k - 1$. Thus, we have the last inequality in (16).

It remains to establish the lower bound $L(\delta)$ and upper bound $U(\delta)$ on $1 - \max\limits_{1}^{k}$ and to show that these bounds are attained, with $\delta = d$, if and only if the a_i's are as in (17) and (18), respectively.

By symmetry, without loss of generality (w.l.o.g.) $a_1 \geqslant \cdots \geqslant a_k$. Then, letting $h_i := a_i - a_{i+1}$ for $i \in [k]$ (with $a_{k+1} := 0$), we have

$$h_1 \geqslant 0, \ldots, h_k \geqslant 0,$$

$$\max\limits_{1}^{k} a_i = a_1 = \sum_{1}^{k} h_i,$$

$$\delta = \sum_{1 \leqslant i < j \leqslant k} (a_i - a_j) = \sum_{1 \leqslant i < j \leqslant k} \sum_{q=i}^{j-1} h_q = \sum_{q=1}^{k-1} h_q \sum_{1 \leqslant i \leqslant q} \sum_{q+1 \leqslant j \leqslant k} 1$$

$$= \sum_{q=1}^{k-1} h_q \, q(k - q) = \sum_{i=1}^{k} i(k - i) h_i,$$

$$1 = \sum_{1}^{k} a_j = \sum_{j=1}^{k} \sum_{i=j}^{k} h_i = \sum_{i=1}^{k} i h_i.$$

Take now indeed any $d \in [0, k - 1]$. Introducing

$$p_i := i h_i$$

for $i \in [k]$, we further restate the conditions on the a_i's (with δ equal the prescribed value $d \in [0, k-1]$, as desired):

$$(19) \qquad p_1 \geqslant 0, \dots, p_k \geqslant 0, \ \sum_{i=1}^{k} p_i = 1,$$

$$(20) \qquad \sum_{i=1}^{k} (k-i)p_i = d \quad \text{or, equivalently,} \quad \sum_{i=1}^{k} ip_i = k - d,$$

and

$$\max_{1}^{k} a_i = a_1 = \sum_{1}^{k} g(i)p_i,$$

where $g(i) := \frac{1}{i}$; here and in the rest of the proof of Lemma 10, i is an arbitrary number in the set $[k]$.

Introduce also

$$g^{U}(i) := g(k-m-1) + [g(k-m) - g(k-m-1)][i - (k-m-1)]$$

and

$$p_i^{U} := (d-m)\,\mathrm{I}\{i = k-m-1\} + (m+1-d)\,\mathrm{I}\{i = k-m\},$$

where

$$m := \lceil d \rceil - 1;$$

here and in the rest of the proof of Lemma 10, i is an arbitrary number in the set $[k]$. One may note at this point that $m \in \{0, \dots, k-2\}$. Note that the function g is strictly convex on the set $[k]$, the function g^{U} is affine, $g^{U} = g$ on the set $\{k-m-1, k-m\}$, and hence $g > g^{U}$ on $[k] \setminus \{k-m-1, k-m\}$. Moreover, conditions (19) and (20) hold with p_i^{U} in place of p_i. So,

$$(21) \quad \sum_{1}^{k} g(i)p_i \geqslant \sum_{1}^{k} g^{U}(i)p_i = g^{U}\Big(\sum_{1}^{k} ip_i\Big)$$

$$= g^{U}(k-d) = g^{U}\Big(\sum_{1}^{k} ip_i^{U}\Big) = \sum_{1}^{k} g^{U}(i)p_i^{U} = \sum_{1}^{k} g(i)p_i^{U};$$

the inequality here holds because $g \geqslant g^{U}$; the first and fourth equalities follow because the function g^{U} is affine; the second and third equalities hold because of the condition (20) for the p_i's and p_i^{U}'s; and the last equality follows because $g^{U}(i) = g(i)$ for i in the set $\{k-m-1, k-m\}$, whereas $p_i^{U} = 0$ for i not in this set. We conclude that, under conditions (19) and (20), $\max_{1}^{k} a_i$ is minimized – or, equivalently, $1 - \max_{1}^{k} a_i$ is maximized – if and only if $p_i = p_i^{U}$ for all i; that is, if and only if $h_i = p_i^{U}/i$ or all i; that is, if and only if the a_i's – related to the p_i's by the formula $a_i = \sum_{j=i}^{k} \frac{1}{i} p_i$ – are as in (18). This concludes the proof of the part of Lemma 10 concerning the upper bound $U(\cdot)$.

The proof of the part of Lemma 10 concerning the lower bound $L(\cdot)$ is similar and even easier. Here let

$$g^{L}(i) := g(1) + [g(k) - g(1)]\frac{i-1}{k-1}$$

and
$$p_i^L := \frac{d}{k-1}\,\mathrm{I}\{i=1\} + \left(1 - \frac{d}{k-1}\right)\mathrm{I}\{i=k\}.$$

Recall that the function g is strictly convex on the set $[k]$. Note that the function g^L is affine, $g^L = g$ on the set $\{1,k\}$, and $g^L > g$ on the set $[k]\setminus\{1,k\}$, so that $g \leqslant g^L$ on $[k]$. Moreover, conditions (19) and (20) hold with p_i^L in place of p_i. So,

$$\sum_1^k g(i)p_i \leqslant \sum_1^k g^L(i)p_i = g^L\Big(\sum_1^k ip_i\Big)$$

$$= g^L(k-d) = g^L\Big(\sum_1^k ip_i^L\Big) = \sum_1^k g^L(i)p_i^L = \sum_1^k g(i)p_i^L;$$

cf. (21). So, under conditions (19) and (20), $\overset{k}{\underset{1}{\max}}\, a_i$ is maximized – or, equivalently, $1 - \overset{k}{\underset{1}{\max}}\, a_i$ is minimized – if and only if $p_i = p_i^L$ for all i; that is, if and only if $h_i = p_i^L/i$ or all i; that is, if and only if the a_i's are as in (17). This concludes the proof of the part of Lemma 10 concerning the upper bound $L(\cdot)$ and hence the entire proof of the lemma. $\qquad\square$

Proof of Proposition 7. We have to show that, under the conditions of this proposition, $L(\Delta) > L_{\mathsf{FM}}(H)$. Recalling (8) and the fact that the function Φ is increasing, we can rewrite inequality $L(\Delta) > L_{\mathsf{FM}}(H)$ as $\Phi(L(\Delta)) > H$. In view of (7), (3), and (5), $H = \ln\ell$, $\Delta = k-\ell$, and $L(\Delta) = \frac{\ell-1}{k}$. So, inequality $\Phi(L(\Delta)) > H$ can be in turn rewritten as

$$(22) \qquad d(\ell) := d_k(\ell) := \Phi\Big(\frac{\ell-1}{k}\Big) - \ln\ell \overset{(?)}{>} 0$$

for k and ℓ as in the conditions in Proposition 7. For $k \in \{6,7,8,9\}$ and $\ell = k-4$, as well as for $k \in \{3,4,5\}$ and $\ell \geqslant 2$ in the set $\{k-3, k-2, k-1\}$, inequality (22) can be verified by direct calculations. So, without loss of generality $k \geqslant 6$ and $\ell \in [k-3, k)$, where we may allow ℓ to take non-integral values as well. Note that

$$d''(\ell)(\ell-1)\ell^2(k+1-\ell) = -1 + 2\ell - 2\ell^2 + (\ell-1)k$$

$$\leqslant -1 + 2\ell - 2\ell^2 + (\ell-1)(\ell+3) = -(\ell-2)^2 < 0$$

for $\ell > 2$. Hence, $d(\ell)$ is strictly concave in $\ell > 2$ such that $\ell \in [k-3, k]$. Also, $d(k) = 0$. So, to complete the proof of Proposition 7, it suffices to show that

$$(23) \qquad \tilde{d}(k) := k\, d_k(k-3)\Big[= (k-4)\ln\frac{k(k-1)}{k-4} + 4\ln\frac{k}{4} - k\ln(k-3)\Big] \overset{(?)}{>} 0$$

for $k \geqslant 6$. We find that

$$\tilde{d}''(k) = -\frac{36}{(k-4)(k-3)^2(k-1)^2 k} < 0$$

for $k \geqslant 6$, and so, $\tilde{d}(k)$ is strictly concave in $k \geqslant 6$. Moreover, $\tilde{d}(6) = 0.446\cdots > 0$ and $\tilde{d}(k) \to 6 - 8\ln 2 = 0.454\cdots > 0$ as $k \to \infty$. Thus, inequality (23) indeed holds for $k \geqslant 6$, and the proof of Proposition 7 is now complete. $\qquad\square$

Proof of Proposition 8. Under the conditions of this proposition,

$$H = -\Big(1 - \frac{\nu-1}{k}\Big)\ln\Big(1 - \frac{\nu-1}{k}\Big) - \frac{\nu-1}{k}\ln\frac{\nu-1}{k(k-1)} \to 0$$

and hence, by (9) and (10), $U_{\mathsf{FM}}(H) \to 0$ as $k \to \infty$.

On the other hand, here $\Delta = k - \nu$. Therefore and because $U(\Delta)$ is decreasing in Δ,

$$U(\Delta) \geqslant U(\lceil \Delta \rceil) = U_{\mathsf{simpl}}(\lceil \Delta \rceil) = 1 - \frac{1}{k - \lceil \Delta \rceil} = 1 - \frac{1}{\lfloor \nu \rfloor},$$

which latter is a positive constant with respect to k and hence does not go to 0 as $k \to \infty$. Thus, the conclusion of Proposition 8 follows. □

Proof of Proposition 9. Here w.l.o.g. $\{\rho_1(x), \rho_2(x)\} = \{1 - p, p\}$ for each $x \in S$, where $p := p_* \in [0, 1/2]$. Then the equalities $U(\Delta) = L(\Delta) = L_{\mathsf{FM}}(H) = p_*$ follow immediately from the definitions.

It remains to show that $U_{\mathsf{FM}}(H) > U_{\mathsf{simpl}}(\Delta) > p$ for $p \in (0, 1/2)$. The second inequality here is obvious, since in this case $U_{\mathsf{simpl}}(\Delta) = \frac{2p}{1+2p}$. To verify that $U_{\mathsf{FM}}(H) > U_{\mathsf{simpl}}(\Delta)$ for $p \in (0, 1/2)$, consider $d(p) := U_{\mathsf{FM}}(H) - U_{\mathsf{simpl}}(\Delta) = -\frac{1}{2}(1-p)\log_2(1-p) - \frac{1}{2}p\log_2 p - \frac{2p}{1+2p}$. It is easy to see that

$$d''(p)(1-p)p(1+2p)^3 \ln 4 = -1 + p(16\ln 2 - 6) - p^2(12 + 16\ln 2) - 8p^3 < 0$$

for $p \in (0, 1/2)$, so that d is strictly concave on $(0, 1/2)$. Also, $d(0+) = d(1/2) = 0$. So, $d > 0$ on $(0, 1/2)$, which completes the proof of Proposition 9. □

In conclusion, let us mention a sample of other related results found in the literature. In [6], for $k = 2$, sharp lower bounds on the misclassification probabilities for three particular classifiers in terms of characteristics generalizing the Kullback–Leibler divergence and the Hellinger distance we obtained. Lower and upper bounds on the misclassification probability based on Renyi's information were given in [2]. Upper and lower bounds on the risk of an empirical risk minimizer for $k = 2$ were obtained in [5] and [4], respectively.

References

[1] D. Erdogmus and J. Principe. Information transfer through classifiers and its relation to probability of error. In *Proceedings. IJCNN '01. International Joint Conference on Neural Networks 2001, (Washington, DC, 2001)*, pages 50–54, 2001.

[2] D. Erdogmus and J. C. Principe. Lower and upper bounds for misclassification probability based on Renyi's information. *Journal of VLSI signal processing systems for signal, image and video technology*, 37:305–317, 2004.

[3] M. Feder and N. Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994.

[4] A. Kontorovich and I. Pinelis. Exact lower bounds for the agnostic probably-approximately-correct (PAC) machine learning model. arXiv:1606.08920 [cs.LG], 2016.

[5] P. Massart and E. Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[6] G. T. Toussaint. On the divergence between two distributions and the probability of misclassification of several decision rules. In *Proceedings of the Second International Joint Conference on Pattern Recognition*, pages 27–34, 1974.

DEPARTMENT OF MATHEMATICAL SCIENCES, MICHIGAN TECHNOLOGICAL UNIVERSITY, HOUGHTON, MICHIGAN 49931, USA, E-MAIL: IPINELIS@MTU.EDU