Bounds on the Bayes Error Given Moments

Bela A. Frigyik and Maya R. Gupta, *Senior Member, IEEE*

***Abstract*—We show how to compute lower bounds for the supremum Bayes error if the class-conditional distributions must satisfy moment constraints, where the supremum is with respect to the unknown class-conditional distributions. Our approach makes use of Curto and Fialkow's solutions for the truncated moment problem. The lower bound shows that the popular Gaussian assumption is not robust in this regard. We also construct an upper bound for the supremum Bayes error by constraining the decision boundary to be linear.**

***Index Terms*—Bayes error, maximum entropy, moment constraint, quadratic discriminant analysis (QDA), truncated moments.**

I. Introduction

A standard approach in pattern recognition is to estimate the first two moments of each class-conditional distribution from training samples, and then assume the unknown distributions are Gaussians. Depending on the exact assumptions, this approach is called *linear* or *quadratic discriminant analysis* (QDA) [1], [2]. Gaussians are known to maximize entropy given the first two moments [3] and to have other nice mathematical properties, but how robust are they with respect to maximizing the Bayes error? To answer that, in this paper, we investigate the more general question: "What is the maximum possible Bayes error given moment constraints on the class-conditional distributions?"

We present both a lower bound and an upper bound for the maximum possible Bayes error. The lower bound means that there exists a set of class-conditional distributions that have the given moments and have a Bayes error above the given lower bound. The upper bound means that no set of class-conditional distributions can exist that have the given moments and have a higher Bayes error than the given upper bound.

Our results provide some insight into how confident one can be in a classifier if one is confident in the estimation of the first n moments. In particular, given only the certainty that two equally likely classes have different means (and no trustworthy estimate of their variances), we show that the Bayes error could be $1/2$, i.e., the classes may not be separable at all. Given the first two moments, our results show that the popular Gaussian assumption for the class distributions is fairly optimistic—the true Bayes error could be much worse. However, we show that the closer the class variances, the more robust the Gaussian assumption. In general, the given lower bound may be a helpful way to assess the robustness of the assumed distributions used in generative classifiers.

The given upper bound may also be useful in practice. Recall that the Bayes error is the error that would arise from the optimal decision boundary. Thus, if one has a classifier and finds that the sample test error is much higher than the given upper bound on the worst case Bayes error, two possibilities should be considered. First, it may imply that the classifier's decision boundary is far from optimal, and that the classifier should be improved. Or, it could be that the test samples used to judge the test error are an unrepresentative set, and that more test samples should be taken to get a useful estimate of the test error.

There are a number of other results regarding the optimization of different functionals given moment constraints (e.g., [4]–[12]). However, we are not aware of any previous work bounding the maximum Bayes error given moment constraints. Some related problems are considered by Antos *et al.* [13]; a key difference to their work is that while we assume moments are given, they instead take as given independent identically distributed (i.i.d.) samples from the class-conditional distributions, and they then bound the average error of an estimate of the Bayes error.

After some mathematical preliminaries, we give lower bounds for the maximum Bayes error in Section III. We construct our lower bounds by creating a truncated moment problem. The existence of a particular lower bound then depends on the feasibility of the corresponding truncated moment problem, which can be checked using Curto and Fialkow's solutions [14] (reviewed in the Appendix). In Section IV, we show that the approach of Lanckreit *et al.* [10], which assumes a linear decision boundary, can be extended to provide an upper bound on the maximum Bayes error. We provide an illustration of the tightness of these bounds in Section V, then end with a discussion and some open questions.

II. Bayes Error

Let $\mathcal{X}$ be a vector space and let $\mathcal{Y}$ be a finite set of classes. Without loss of generality, we may assume that $\mathcal{Y} = \{1, \ldots, G\}$. Suppose that there is a measurable classification function $h : \mathcal{X} \to S_G$, where S_G is the $(G-1)$ probability simplex. Then, the ith component of $h(x)$ can be interpreted as the probability of class i given x, and we write $p(i|x) = h(x)_i$.

For a given $x \in \mathcal{X}$, the Bayes classifier selects the class $\hat{y}(x)$ that maximizes the posterior probability $p(i|x)$ (if there is a tie for the maximum, then any of the tied classes can be chosen).

Manuscript received April 15, 2011; revised November 06, 2011, December 01, 2011, and January 15, 2012; accepted January 25, 2012. Date of publication February 13, 2012; date of current version May 15, 2012. This work was supported by the United States PECASE Award.

The authors are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: frigyik@uw.edu; gupta@ee.washington.edu).

Communicated by N. Cesa-Bianchi, Associate Editor for Pattern Recognition, Statistical Learning, and Inference.

Digital Object Identifier 10.1109/TIT.2012.2187634

0018-9448/$31.00 © 2012 IEEE

The probability that the Bayes classifier is wrong for a given x is

$$P_e(x) = 1 - \max_{i\in\mathcal{Y}} p(i|x).$$

Suppose there is a probability measure ν defined on $\mathcal{X}$. Then, the *Bayes error* is the expectation of P_e:

$$\mathbb{E}[P_e] = 1 - \int_{x\in\mathcal{X}} \max_{i\in\mathcal{Y}} p(i|x) d\nu(x). \tag{1}$$

The fact that the $p(i|x)$ must sum to one over i, and thus $\max_i p(i|x) \geq 1/G$, implies a trivial upper bound on the Bayes error given in (1)

$$\begin{aligned}\mathbb{E}[P_e] &\leq 1 - \int_x \frac{1}{G} d\nu(x)\\ &= \frac{G-1}{G}.\end{aligned}$$

Suppose that the probability measure ν is defined on $\mathcal{X}$ such that it is absolutely continuous w.r.t. the Lebesgue measure such that it has density $p(x)$. Or suppose that it is discrete and expressed as

$$\nu = \sum_{j=1}^{\infty} \alpha_j \delta_{x_j}$$

where δ_{x_j} is the Dirac measure with support x_j, $\alpha_j > 0$ for all $j = 1, 2, \ldots$ and $\sum_{j=1}^{\infty} \alpha_j = 1$, and we say the density $p(x_j) = \alpha_j$. In either case, (1) can be expressed in terms of the ith class prior $p(i) = \int p(i|x) d\nu(x)$ and ith class-conditional density $p(x|i)$ (or probability mass function $p(x_j|i)$) as follows:

$$\mathbb{E}[P_e] = \begin{cases} 1 - \sum_{j=1}^{\infty} \max_{i\in\mathcal{Y}} p(x_j|i)p(i), & \text{discrete case} \\ 1 - \int \max_{i\in\mathcal{Y}} p(x|i)p(i)dx, & \text{abs. continuous case.} \end{cases} \tag{2}$$

If ν is a general measure, then Lebesgue's decomposition theorem says that it can be written as a sum of three measures: $\nu = \nu_d + \nu_{ac} + \nu_{sc}$. Here, ν_d is a discrete measure and the other two measures are continuous, ν_{ac} is absolutely continuous w.r.t. the Lebesgue measure, and ν_{sc} is the remaining singular part. We have a convenient representation for both the discrete and the absolutely continuous part of a measure but not for the singular portion. For this reason, we are going to restrict our attention to measures that are either discrete or absolutely continuous (or a linear combination of these kind of measures).

III. Lower Bounds for Worst Case Bayes Error

Our strategy to providing a lower bound on the supremum Bayes error is to constrain the G probability distributions $p(x|i)$, $i \in \mathcal{Y}$ to have an overlap of size $\epsilon \in (0, 1)$. Specifically, we constrain the G distributions to each have a Dirac measure of size ϵ at the same location. In the case of uniform class prior probabilities, this makes the Bayes error at least $\epsilon\frac{G-1}{G}$. The largest such ϵ for which this overlap constraint is feasible determines the best lower bound on the worst case Bayes error this strategy can provide. The maximum such feasible ϵ can be determined by checking whether there is a solution to a corresponding truncated moment problem (see the Appendix for details). Note that this approach does not restrict the distributions from overlapping elsewhere which would increase the Bayes error, and thus this approach only provides a lower bound to the maximum Bayes error.

We first present a constructive solution showing that no matter what the first moments are, the Bayes error can be arbitrarily bad if only the first moments are given. Then, we derive conditions for the size of the lower bound for the two moment case and three moment case, and end with what we can say for the general case of n moments.

Lemma 1:

Suppose the first moments $\gamma_{1,i}$ are given for each i in a subset of $\{1, \ldots, G\}$ and the remaining class-conditional distributions are unconstrained. Then, for all $1 > \epsilon > 0$, one can construct G discrete or absolutely continuous class-conditional distributions such that the Bayes error $\mathbb{E}[P_e] \geq (1 - \max_{i\in\mathcal{Y}} p(i))\epsilon$.

Proof: This lemma works for any vector space $\mathcal{X}$. The moment constraints hold if the ith class-conditional distribution is taken to be $p(x|i) = \epsilon\delta_0(x) + (1-\epsilon)\delta_{z_i}(x)$ where $z_i = \frac{\gamma_{1,i}}{(1-\epsilon)}$. This constructive solution exists for any $\epsilon \in (0, 1)$ and yields a Bayes error of at least $(1 - \max_{i\in\mathcal{Y}} p(i))\epsilon$. To see this, substitute the G measures $\{\epsilon\delta_0 + (1-\epsilon)\delta_{z_i}\}$ into the discrete case of (2) to produce

$$\begin{aligned}\mathbb{E}[P_e] &= 1 - \sum_{j=1}^{\infty} \max_{i\in\mathcal{Y}} \left(\epsilon\delta_0(x_j) + (1-\epsilon)\delta_{z_i}(x_j)\right) p(i)\\ &= 1 - \left(\max_{i\in\mathcal{Y}} p(i)\epsilon + \sum_{j=1}^{\infty} \max_{i\in\mathcal{Y}} p(i)(1-\epsilon)\delta_{z_i}(x_j)\right)\\ &\geq 1 - \left(\epsilon \max_{i\in\mathcal{Y}} p(i) + (1-\epsilon)\sum_{i=1}^{G} p(i)\right)\\ &= \epsilon\left(1 - \max_{i\in\mathcal{Y}} p(i)\right).\end{aligned}$$

For an absolutely continuous example, consider $\mathcal{X} = \mathbb{R}$. The uniform densities $p_l(x|i) = \frac{1}{2lp(i)}\mathbb{1}_{[\gamma_{1,i}-(lp(i)),\gamma_{1,i}+(lp(i))]}(x)$ with $i = 1, 2, \ldots, G$ (where $\mathbb{1}_E$ is the indicator function of the set E) provide class-conditional distributions such that as $l \to \infty$ the Bayes error goes to $1 - \max_{i\in\mathcal{Y}} p(i)$. To see this, let $i^* \in \arg\max_{i\in\mathcal{Y}} p(i)$ and consider the difference $d_i = \gamma_{1,i^*} + lp(i^*) - (\gamma_{1,i} + lp(i)) = (\gamma_{1,i^*} - \gamma_{1,i}) + l(p(i^*) - p(i))$. If $p(i^*) - p(i) > 0$, then $d_i \to \infty$ as $l \to \infty$; therefore, there is an $l' > 0$ such that if $l > l'$, then $d_i > 0$, and hence, $\gamma_{1,i^*} + lp(i^*) > \gamma_{1,i} + lp(i)$. A similar derivation shows that there is an $l'' > 0$ such that if $l > l''$, then $\gamma_{1,i^*} - lp(i^*) < \gamma_{1,i} - lp(i)$. In other words, if $p(i^*) - p(i) > 0$, then $p(i^*)p(x|i^*)$ eventually dominates $p(i)p(x|i)$ since all the functions $p(j)p(x|j)$, $j = 1, \ldots, G$ have the same amplitude $\frac{1}{2l}$. If $p(i^*) - p(i) = 0$, then the integral of the function $p(i)p(x|i)$ that is not dominated by $p(i^*)p(x|i^*)$ is $\frac{|\gamma_{1,i^*} - \gamma_{1,i}|}{2l} \to 0$ as $l \to \infty$. Finally, the integral of the dominant function $p(i^*)p(x|i^*)$ is $\frac{1}{2l}2lp(i^*) = p(i^*)$,

and therefore, the Bayes error approaches $1 - \max_{i\in\mathcal{Y}} p(i)$ as $l \to \infty$. ■

Theorem 1:

Suppose that $\mathcal{X} = \mathbb{R}$ and that there exist G class-conditional measures with first and second moments $\{\gamma_{1,i}, \gamma_{2,i}\}, i \in \mathcal{Y}$. Given only this set of moments $\{\gamma_{1,i}, \gamma_{2,i}\}$, $i \in \mathcal{Y}$, a lower bound on the supremum Bayes error is

$$\begin{aligned}\sup \mathbb{E}[P_e] &\geq \sup_{\Delta\in\mathbb{R}} \left[\sum_{i=1}^{G} p(i) \left(1 - \frac{(\gamma_{1,i} - \Delta)^2}{\gamma_{2,i} + \Delta^2 - 2\Delta\gamma_{1,i}}\right) \right. \\ &\quad \left. - \max_{i\in\mathcal{Y}} \left\{ p(i) \left(1 - \frac{(\gamma_{1,i} - \Delta)^2}{\gamma_{2,i} + \Delta^2 - 2\Delta\gamma_{1,i}}\right)\right\}\right] \\ &\geq (G-1) \sup_{\Delta\in\mathbb{R}} \min_{i\in\mathcal{Y}} \left\{ p(i) \left(1 - \frac{(\gamma_{1,i} - \Delta)^2}{\gamma_{2,i} + \Delta^2 - 2\Delta\gamma_{1,i}}\right)\right\},\end{aligned}$$

where the supremum on the left-hand side is taken over all combinations of G class-conditional measures satisfying the moment constraints.

If $G = 2$, then

$$\sup \mathbb{E}[P_e] \geq \sup_{\Delta\in\mathbb{R}} \min_{i=1,2} \left\{ p(i) \left(1 - \frac{(\gamma_{1,i} - \Delta)^2}{\gamma_{2,i} + \Delta^2 - 2\Delta\gamma_{1,i}}\right)\right\}.$$

Further, if the class priors are equal, then, in terms of the centered second moment $\sigma_i^2 = \gamma_{2,i} - \gamma_{1,i}^2$, the optimal Δ value is one of

$$\Delta = \frac{-(\gamma_{1,2}\sigma_1^2 - \gamma_{1,1}\sigma_2^2) \pm \sigma_1\sigma_2 \left|\gamma_{1,1} - \gamma_{1,2}\right|}{\sigma_2^2 - \sigma_1^2} \tag{3}$$

if $\sigma_1 \neq \sigma_2$. Otherwise, if $\sigma_1 = \sigma_2$

$$\Delta = \frac{\gamma_{1,1} + \gamma_{1,2}}{2} \tag{4}$$

and the lower bound simplifies

$$\sup \mathbb{E}[P_e] \geq \frac{2\sigma_1^2}{4\sigma_1^2 + (\gamma_{1,1} - \gamma_{1,2})^2}.$$

Proof: Consider some $0 < \epsilon < 1$. If the class prior is uniform, then a sufficient condition for the Bayes error to be at least $\frac{G-1}{G}\epsilon$ is if all of the unknown measures share a Dirac measure of at least ϵ. First, we place this Dirac measure at zero and find the maximum ϵ for which this can be done. Then, later in the proof, we show that a larger ϵ (and hence a tighter lower bound on the maximum Bayes error) can be found by placing this shared Dirac measure in a more optimal location, or equivalently, by shifting all the measures.

Suppose a probability measure μ can be expressed in the form $\epsilon\delta_0 + \tilde{\mu}$ where $\tilde{\mu}$ is some measure such that $\tilde{\mu}(\{0\}) = 0$. If μ satisfies the original moment constraints, then $\tilde{\mu}$ also satisfies them; this follows directly from the moment definition for $n \geq 1$:

$$\int x^n d\mu(x) = 0^n\epsilon + \int x^n d\tilde{\mu}(x) = \int x^n d\tilde{\mu}(x).$$

Also $\tilde{\mu}(\mathcal{X}) = 1 - \epsilon$. Thus, we require a measure $\tilde{\mu}$ with a zeroth moment $\gamma_0 = 1 - \epsilon > 0$ and the original first and second moments γ_1, γ_2. Then, as described in the Appendix, there are two conditions that we have to check. In order to have a measure with the prescribed moments, the matrix

$$A = A(1) = \begin{bmatrix} 1-\epsilon & \gamma_1 \\ \gamma_1 & \gamma_2 \end{bmatrix}$$

has to be positive semidefinite, which holds if and only if $\epsilon \leq 1 - \frac{\gamma_1^2}{\gamma_2}$. (Note that the Theorem assumes that there exists a distribution with the given moments, and thus the above implies that $\gamma_2 \geq \gamma_1^2$.) Moreover, the rank of matrix A and the rank of γ (for notation, see the Appendix) have to be the same. Matrix A can have rank 1 or 2. If $\mathrm{rank}(A) = 1$, then the columns of A are linearly dependent, and therefore, $\mathrm{rank}(\gamma) = 1$. If $\mathrm{rank}(A) = 2$, then A is invertible and $\mathrm{rank}(\gamma) = 2$. Thus, there is a measure $\tilde{\mu}$ with moments $\{1-\epsilon, \gamma_1, \gamma_2\}$ iff $0 \leq \epsilon \leq 1 - \frac{\gamma_1^2}{\gamma_2}$. If such a $\tilde{\mu}$ exists, then there also exists a discrete probability measure with moments $\{1-\epsilon, \gamma_1, \gamma_2\}$ and $0 \leq \epsilon \leq 1 - \frac{\gamma_1^2}{\gamma_2}$ by Curto and Fialkow's results [14, Ths. 3.1 and 3.9].

Suppose we have G such discrete probability measures satisfying the corresponding moments constraints given in the statement of this theorem. Denote the ith discrete probability measure by $\nu_i = \epsilon_i\delta_0 + \sum_{j=1}^{\infty} \alpha_{j,i}\delta_{x_j}$ where $0 \leq \epsilon_i \leq 1 - \frac{\gamma_{1,i}^2}{\gamma_{2,i}}$ and $x_j \neq 0$ for all j, and j indexes the set of all non-zero atoms in the G discrete measures $\{\nu_i\}$. Then, the supremum Bayes error is bounded below by the Bayes error for this set of discrete measures:

$$\begin{aligned}\sup \mathbb{E}[P_e] &\geq 1 - \max_{i\in\mathcal{Y}}\{\epsilon_i p(i)\} - \sum_{j=1}^{\infty} \max_{i\in\mathcal{Y}} \alpha_{j,i} p(i) \\ &\geq 1 - \max_{i\in\mathcal{Y}}\{\epsilon_i p(i)\} - \sum_{j=1}^{\infty}\sum_{i=1}^{G} \alpha_{j,i} p(i) \\ &= 1 - \max_{i\in\mathcal{Y}}\{\epsilon_i p(i)\} - \sum_{i=1}^{G} p(i) \sum_{j=1}^{\infty} \alpha_{j,i} \\ &= 1 - \max_{i\in\mathcal{Y}}\{\epsilon_i p(i)\} - \sum_{i=1}^{G} p(i)(1-\epsilon_i) \\ &= \sum_{i=1}^{G} p(i)\epsilon_i - \max_{i\in\mathcal{Y}}\{\epsilon_i p(i)\}.\end{aligned} \tag{5}$$

This is true for any collection of ϵ_i, $\epsilon_i \leq 1 - \frac{\gamma_{1,i}^2}{\gamma_{2,i}}$. This means that $\sup \mathbb{E}[P_e]$ is an upper bound for (5) for these admissible ϵ_i and we can find a tighter inequality by finding the supremum of (5) over the set of admissible ϵ_i. The domain of the function (5) is the Cartesian product $\prod_{i=1}^{G} \left[0, 1 - \frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right]$. It is a nonempty compact set and (5) is continuous, so we can expect to find a maximum. The maximum is unique and to find it let $(\epsilon_1, \ldots, \epsilon_G)$ be any element in the domain and let $i^* \in \arg\max_i \left\{p(i)\left(1 - \frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right)\right\}$. Since $p(i) \geq 0$, we have

$$\begin{aligned}\sum_{i=1}^{G} p(i)\epsilon_i - \max_{i\in\mathcal{Y}}\{p(i)\epsilon_i\} &\le \sum_{i\neq i^*} p(i)\epsilon_i \\ &\le \sum_{i\neq i^*} p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right) \\ &= \sum_{i=1}^{G} p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right) \\ &\quad - \max_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right)\right\}.\end{aligned}$$

Therefore

$$\begin{aligned}\sup \mathbb{E}[P_e] &\ge \sum_{i=1}^{G} p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right) - \max_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right)\right\} \\ &\ge (G-1)\min_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right)\right\} \qquad (6)\end{aligned}$$

where the supremum on the left-hand side is taken over all the combination of class-conditional measures that satisfy the given moments constraints.

The next step follows from the fact that the Lebesgue measure and the counting measure are shift-invariant measures and the Bayes error is computed by integrating some functions against those measures. Suppose we had G class distributions, and we shift each of them by Δ. The Bayes error would not change. However, our lower bound given in (6) depends on the actual given means $\{\gamma_{1,i}\}$, and in some cases we can produce a better lower bound by shifting the distributions before applying the above lower bounding strategy. The shifting approach we present next is equivalent to placing the shared ϵ measure someplace other than at the origin.

Shifting a distribution by Δ does change all of the moments (because they are not centered moments), specifically, if μ is a probability measure with finite moments $\gamma_0 = 1, \gamma_1, \ldots, \gamma_n$, and μ_Δ is the measure defined by $\mu_\Delta(D) = \mu(D+\Delta)$ for all μ-measurable sets D, then the nth noncentered moment of the shifted measure μ_Δ is

$$\begin{aligned}\tilde{\gamma}_n = \int x^n d\mu_\Delta(x) &= \int (x-\Delta)^n d\mu(x) \\ &= \sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\Delta^{n-k}\gamma_k\end{aligned}$$

where the second equality can easily be proven for any σ-finite measure using the definition of integral. This same formula shows that shifting back the measure will transform back the moments.

For the two-moment case, the shifted measure's moments are related to the original moments by

$$\begin{aligned}\tilde{\gamma}_1 &= \gamma_1 - \Delta \\ \tilde{\gamma}_2 &= \gamma_2 + \Delta^2 - 2\Delta\gamma_1.\end{aligned}$$

Then, a tighter lower bound can be produced by choosing the shift Δ that maximizes the shift-dependent lower bound given in (6):

$$\begin{aligned}&\sup \mathbb{E}[P_e] \\ &\quad \ge \sup_{\Delta\in\mathbb{R}}\left[\sum_{i=1}^{G} p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}\right)\right. \\ &\qquad \left. - \max_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}\right)\right\}\right] \\ &\quad \ge (G-1)\sup_{\Delta\in\mathbb{R}}\min_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}\right)\right\}.\end{aligned}$$

If $G = 2$, then this lower bound is

$$\begin{aligned}&\sup_{\Delta\in\mathbb{R}}\left[\sum_{i=1}^{2} p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}\right)\right. \\ &\qquad \left. - \max_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}\right)\right\}\right] \\ &\quad = \sup_{\Delta\in\mathbb{R}}\min_{i=1,2}\left\{p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}\right)\right\}.\end{aligned}$$

We can say more in the case of equal class priors, that is, if $p(1) = p(2) = 1/2$. The functions

$$f_i(\Delta) = 1 - \frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}$$

are maximized at $\Delta = \gamma_{1,i}$, where the maximum value is 1 and the derivative of function f_i is strictly positive for $\Delta < \gamma_{1,i}$ and strictly negative for $\Delta > \gamma_{1,i}, i = 1, 2$. This means that the potential maximum occurs at the point where the two functions are equal. This results in a quadratic equation if $\sigma_2^2 \neq \sigma_1^2$ with solutions (3), and otherwise a linear one with solution (4).

If $\gamma_{1,1} = \gamma_{1,2}$, then the function f_i with smaller $\gamma_{2,i}$ will provide us with the lower bound which is $1/2$, as expected. If $\gamma_{1,1} \neq \gamma_{1,2}$, then since $f_i(\gamma_{1,i}) = 1$ the maximum occurs at a Δ value which is between the two $\gamma_{1,i}$. To see this, let J be the interval defined by the two $\gamma_{1,i}$. As a consequence of the strict nature of the derivatives, for any Δ value outside of the interval J, the function

$$\min_{i\in\mathcal{Y}}\left\{1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}\right\}$$

is less than on J. But on J, the function $f_1(\Delta) - f_2(\Delta)$ is continuous and thanks to the fact that $f_i(\gamma_{1,i}) = 1$ and the behavior of the derivatives, it has different sign at the two endpoints of J. This means that there is a $\Delta \in J$ such that $f_1(\Delta) - f_2(\Delta) = 0$. ■

This theorem applies only to 1-D distributions. The approach of constraining the distributions to have measure ϵ at a common location can be extended to higher-dimensions, but actually determining whether the moment constraints can still be satisfied becomes significantly hairier; see [14] for a sketch of the truncated moment solutions for higher dimensions.

An argument similar to the one given in the last two paragraphs of the previous proof can be used to show that if $\gamma_{2,i}-\gamma_{1,i}^2$ are all equal for all $i \in \mathcal{Y}$ and any finite G, and if the class priors are equal, then the optimal Δ is $\frac{\gamma_{1,\min}+\gamma_{1,\max}}{2}$, where $\gamma_{1,\min}$ and $\gamma_{1,\max}$ are the smallest and the largest values in the set $\{\gamma_{1,i}\}_{i\in\mathcal{Y}}$, respectively. To see this, we start with rewriting the function f_i:

$$f_i(\Delta) = 1 - \frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}} = \frac{\sigma_i^2}{\sigma_i^2+(\Delta-\gamma_{1,i})^2}.$$

This shows that if the condition mentioned previously holds, then the functions f_i are shifted versions of each other. Let $f_{\min}$ and $f_{\max}$ be the functions corresponding to $\gamma_{1,\min}$ and $\gamma_{1,\max}$, respectively, and let Δ' be the point where $f_{\min}$ and $f_{\max}$ intersect. Because of the strict nature of the derivatives of f_i and because the functions are shifted versions of each other, for any $\Delta \geq \Delta'$, $f_{\min}$ is smaller than any other f_i. Because of symmetry, it is true that for any $\Delta \leq \Delta'$, $f_{\max}$ is smaller than any other f_i. Again, by symmetry, we have that $\Delta' = \frac{\gamma_{1,\min}+\gamma_{1,\max}}{2}$, and therefore, this is the optimal Δ.

Corollary 1: Suppose that $\mathcal{X} = \mathbb{R}$ and that the first, the second, and the third moments are given for G class-conditional measures, i.e., for the ith class-conditional measure we are given $\{\gamma_{1,i},\gamma_{2,i},\gamma_{3,i}\}$. Then, the Bayes error has lower bound

$$\begin{aligned}
&\sup \mathbb{E}[P_e] \\
&\quad \geq \sup_{\delta>0}\sup_{\Delta\in\mathbb{R}}\left[\sum_{i=1}^{G} p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}-\delta\right)\right. \\
&\qquad \left. -\max_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}-\delta\right)\right\}\right] \\
&\quad \geq (G-1)\sup_{\Delta\in\mathbb{R}}\min_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{(\gamma_{1,i}-\Delta)^2}{\gamma_{2,i}+\Delta^2-2\Delta\gamma_{1,i}}\right)\right\}.
\end{aligned}$$

Proof: In this case, we have a list of four numbers $\{1-\epsilon,\gamma_1,\gamma_2,\gamma_3\}$ and again

$$A = A(1) = \begin{bmatrix} 1-\epsilon & \gamma_1 \\ \gamma_1 & \gamma_2 \end{bmatrix}.$$

If $\gamma_0 = 1-\epsilon > 0$, then A is positive definite if $\epsilon \leq 1-\frac{\gamma_1^2}{\gamma_2}-\delta < 1-\frac{\gamma_1^2}{\gamma_2}$. In this case, $\mathbf{v}_2 = (\gamma_2,\gamma_3)^T$ and it is in the range of A since A is invertible. The statements in the Appendix imply that there is a measure with moments $\{1-\epsilon,\gamma_1,\gamma_2,\gamma_3\}$ and consequently that

$$\begin{aligned}
\sup \mathbb{E}[P_e] &\geq \sup_{\delta>0}\left[\sum_{i=1}^{G} p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}-\delta\right)\right. \\
&\qquad \left. -\max_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}-\delta\right)\right\}\right] \\
&\geq (G-1)\min_{i\in\mathcal{Y}}\left\{p(i)\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right)\right\}.
\end{aligned}$$

The rest of the proof follows analogously to the proof of Theorem 1. ■

The proof of Corollary 1 relies on the fact that for $\delta > 0$ the matrix $A(1)$ featured in the proof is invertible, so one of the conditions for the existence of a measure with the given moments is automatically satisfied (see Appendix). If $\delta = 0$ then $A(1)$ is only positive semidefinite and it is not obvious that the vector $\mathbf{v}_2$ is in the range of $A(1)$.

The following lemma is stated for completeness.

Lemma 2: Suppose that $\mathcal{X} = \mathbb{R}$ and that the first n moments are given for G equally likely class-conditional measures, i.e., for the ith class-conditional measure, we are given $\{\gamma_{1,i},\gamma_{2,i},\ldots,\gamma_{n,i}\}$. Then, if there exist measures of the form $\epsilon_i\delta_0+\nu_i$ where ν_i satisfies the moments conditions given above the corresponding Bayes error can be bounded from below:

$$\begin{aligned}
\sup \mathbb{E}[P_e] &\geq 1-\frac{1}{G}\left[\max_{i\in\mathcal{Y}}\{\epsilon_i\}+\sum_{i=1}^{G}(1-\epsilon_i)\right] \\
&= \frac{1}{G}\left[\sum_{i=1}^{G}\epsilon_i-\max_{i\in\mathcal{Y}}\{\epsilon_i\}\right]
\end{aligned}$$

where the supremum on the left is taken over all the measures satisfying the moment constraints noted above.

Proof: The first part of the proof of Theorem 1 is applicable in this case. ■

As in the case for two moments, the lower bound can be further tightened by optimizing over all possible shifts of the overlap Dirac measure.

IV. Upper Bound for Worst Case Bayes Error

Because the Bayes error is the smallest error over all decision boundaries, one approach to constructing an upper bound on the worst case Bayes error is to restrict the set of considered decision boundaries to a set for which the worst case error is easier to analyze. In related work, Lanckreit *et al.* [10] take as given the first and second moments of each class-conditional distribution, and attempt to find the linear decision boundary classifier that minimizes the worst case classification error rate with respect to any choice of class-conditional distributions that satisfy the given moment constraints. Here, we show that this approach can be extended to produce an upper bound on the supremum Bayes error for the $G = 2$ case.

Let $\mathcal{X}$ be any feature space. Suppose one has two fixed class-conditional measures ν_1,ν_2 on $\mathcal{X}$. As in Lanckreit *et al.* [10], consider the set of linear decision boundaries. Any linear decision boundary splits the domain into two half-spaces S_1 and S_2. We work with linear decision boundaries because these are the only kind of decision boundaries that split the domain into two convex subsets. The error produced by a linear decision boundary corresponding to the split (S_1,S_2) is

$$p(1)\nu_1(S_2)+p(2)\nu_2(S_1) \geq \mathbb{E}[P_e](\nu_1,\nu_2).$$

That is, the error from any linear decision boundary upper bounds the Bayes error $\mathbb{E}[P_e]$ for two given measures. To obtain a tighter upper bound on the Bayes error, minimize the left-hand side over all linear decision boundaries

$$\inf_{S_1,S_2}\left(p(1)\nu_1(S_2)+p(2)\nu_2(S_1)\right) \geq \mathbb{E}[P_e](\nu_1,\nu_2).$$

Now suppose ν_1 and ν_2 are unknown, but their first moments (means) and second centered moments (μ_1,Σ_1) and (μ_2,Σ_2)

are given. Then, we note that the supremum (over all measures ν_1 and ν_2 with those moments) of the smallest linear decision boundary error forms an upper bound on the supremum Bayes error, where the supremum is taken with respect to the feasible measures ν_1, ν_2:

$$\begin{aligned}\sup \mathbb{E}[P_e] &\leq \sup_{\nu_1|\mu_1,\Sigma_1} \sup_{\nu_2|\mu_2,\Sigma_2} \inf_{S_1,S_2} \left(p(1)\nu_1(S_2) + p(2)\nu_2(S_1)\right) \\ &\leq \inf_{S_1,S_2} \left(p(1) \sup_{\nu_1|\mu_1,\Sigma_1} \nu_1(S_2) + p(2) \sup_{\nu_2|\mu_2,\Sigma_2} \nu_2(S_1) \right).\end{aligned} \tag{7}$$

This upper bound can be simplified using the following result[1] by Bertsimas and Popescu [15] (which follows from a result by Marshall and Olkin [16]):

$$\sup_{\nu|\mu,\Sigma} \nu(S) = \frac{1}{1+c(S)}, \quad \text{where } c(S) = \inf_{x\in S} (x-\mu)^T \Sigma^{-1} (x-\mu) \tag{8}$$

where the sup in (8) is over all probability measures ν with domain $\mathcal{X}$, mean μ, and centered second moment Σ; and S is any convex set in the domain of ν.

Since S_1 and S_2 in (7) are half-spaces, they are convex and (8) can be used to quantify the upper bound. To simplify this upper bound, for the rest of this section, let $\mathcal{X}$ be 1-D. Then, the covariance matrices Σ_1 and Σ_2 are just scalars that we denote by σ_1^2 and σ_2^2, respectively. In one dimension, any decision boundary that results in a half-plane split is simply a point $s \in \mathbb{R}$. Without loss of generality with respect to the Bayes error, let $\mu_1 = 0$ and $\mu_1 \leq \mu_2$. Then, $c(S_1)$ and $c(S_2)$ in (8) simplify (for details, see [10, Appendix A]), so that (7) becomes

$$\sup \mathbb{E}[P_e] \leq \min \left\{ \inf_s \left(\frac{p(1)}{1+\frac{s^2}{\sigma_1^2}} + \frac{p(2)}{1+\frac{(\mu_2-s)^2}{\sigma_2^2}} \right), \frac{1}{2} \right\}. \tag{9}$$

If $\sigma_1 = \sigma_2$ and $p(1) = p(2) = 1/2$, then the infimum occurs at $s = \mu_2/2$ and the upper bound becomes

$$\sup \mathbb{E}[P_e] \leq \min \left\{ \frac{4}{4+\frac{\mu_2^2}{\sigma_1^2}}, \frac{1}{2} \right\}.$$

For this case, the given upper bound is twice the given lower bound.

V. Comparison to Error With Gaussians

We illustrate the bounds described in this paper for the common case that the first two moments are known for each class, and the classes are equally likely. We compare with the Bayes error produced under the assumption that the distributions are Gaussians with the given moments. In both cases the first distribution's mean is 0 and the variance is 1, and the second distribution's mean is varied from 0 to 25 as shown on the *x*-axis. The second distribution's variance is 1 for the comparison shown in the top of Fig. 1. The second distribution's variance is 5 for the comparison shown in the bottom of Fig. 1. For the first case, $\sigma_1 = \sigma_2$ so the infimum in (9) occurs at $s = \mu_2/2$ and the upper bound is $\min\{4/(4+\mu_2^2), \frac{1}{2}\}$. For the second case with different variances we compute (9) numerically.

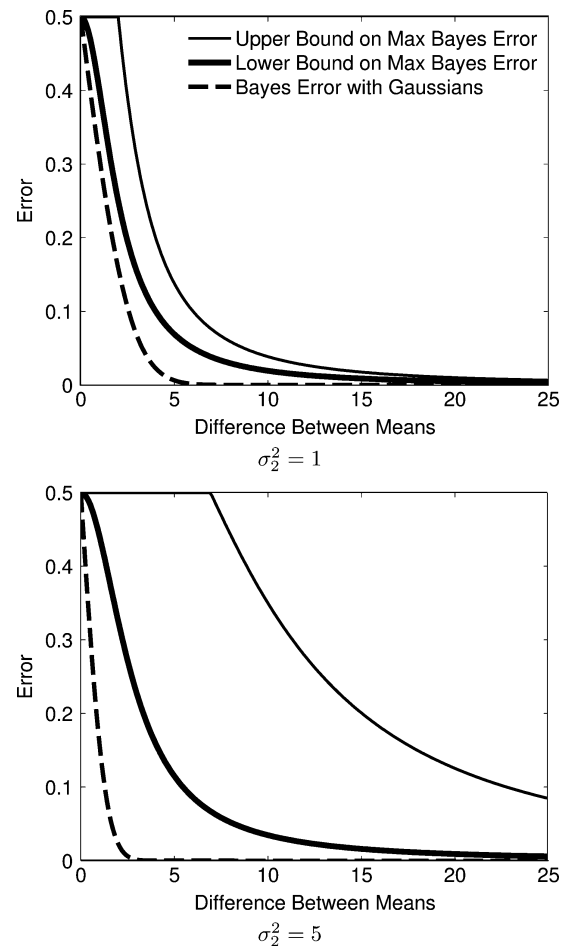

Fig. 1. Comparison of the given lower bound for the worst case Bayes error with the Bayes error produced by Gaussian class-conditional distributions.

Fig. 1 shows that the Bayes error produced by the Gaussian assumption is optimistic compared to the given lower bound for the worst case (maximum) Bayes error. Further, the difference between the Gaussian Bayes error and the lower bound is much larger in the second case when the variances of the two distributions differ.

VI. Discussion and Some Open Questions

We have provided a lower and upper bound on the worst case Bayes error, but a number of open questions arise from this study.

Lower bounds for the worst case Bayes error can be constructed by constraining the distributions. We have shown that constraining the distributions to be Gaussians produces a weak lower bound, and we provided a tighter lower bound by constraining the distributions to overlap in a Dirac measure of ϵ. Given only first moments, our lower bound is tight in that it is arbitrarily close to the worst possible Bayes error. Given two moments, we have shown that the common QDA Gaussian assumption for class-conditional distributions is much more optimistic than our lower bound and increasingly optimistic for increased difference between the variances. However, because

[1]Some readers may recognize this result as a strengthened and generalized version of the Chebyshev–Cantelli inequality.

in our constructions we do not control all the possible overlap between the class-conditional distributions, we believe it should be possible to construct tighter lower bounds.

On the other hand, upper bounds on the worst case Bayes error can be constructed by constraining the considered decision boundaries. Here, we considered an upper bound resulting from restricting the decision boundary to be linear. For the two moment case, we have shown that work by Lanckreit *et al.* leads almost directly to an upper bound. However, the inequality we had to introduce in (7) when we switched the inf and sup may make this upper bound loose. It remains an open question if there are conditions under which the upper bound is tight.

Our result that the popular Gaussian assumption is generally not very robust in terms of worst case Bayes error prompts us to question whether there are other distributions that are mathematically or computationally convenient to use in generative classifiers that would have a Bayes error closer to the given lower bound.

In practice, a moment constraint is often created by estimating the moment from samples drawn i.i.d. from the distribution. In that case, the moment constraint need not be treated as a hard constraint as we have done here. Rather, the observed samples can imply a probability distribution over the moments, which in turn could imply a distribution over corresponding bounds on the Bayes error. A similar open question is a sensitivity analysis of how changes in the moments would affect the bounds.

Lastly, consider the opposite problem: given constraints on the first n moments for each of the class-conditional distributions, how small could the Bayes error be? It is tempting to suppose that one could generally find discrete measures that overlapped nowhere, such that the Bayes error was zero. However, the set of measures that satisfy a set of moment constrains may be nowhere dense, and that impedes us from being able to make such a guarantee. Thus, this remains an open question.

Appendix
Existence of Measures With Certain Moments

The proof of our theorem reduces to the problem of how to check if a given list of n numbers could be the moments of some measure. This problem is called the truncated moment problem; here, we review the relevant solutions by Curto and Fialkow [14].

Suppose we are given a list of numbers $\gamma = \{\gamma_0, \gamma_1, \ldots, \gamma_n\}$, with $\gamma_0 > 0$. Can this collection be a list of moments for some positive Borel measure ν on $\mathbb{R}$ such that

$$\gamma_i = \int s^i d\nu(s)? \tag{10}$$

Let $k = \lfloor n/2 \rfloor$, and construct a Hankel matrix $A(k)$ from γ where the ith row of A is $[\gamma_{i-1}\gamma_i \ldots \gamma_{i-1+k}]$. For example, for $n = 2$ or $n = 3$, $k = 1$:

$$A(1) = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_2 \end{bmatrix}.$$

Let $\mathbf{v}_j$ be the transpose of the vector $(\gamma_{i+j})_{i=0}^k$. For $0 \le j \le k$ this vector is the $j+1$th column of $A(k)$. Define $\mathrm{rank}(\gamma) = k+1$ if $A(k)$ is invertible, and otherwise $\mathrm{rank}(\gamma)$ is the smallest r such that $\mathbf{v}_r$ is a linear combination of $\{\mathbf{v}_0, \ldots, \mathbf{v}_{r-1}\}$.

Then, whether there exists a ν that satisfies (10) depends on n and k:

1) If $n = 2k+1$, then there exists such a solution ν if $A(k)$ is positive semidefinite and $\mathbf{v}_{k+1}$ is in the range of $A(k)$.
2) If $n = 2k$, then there exists such a solution ν if $A(k)$ is positive semidefinite and $\mathrm{rank}(\gamma) = \mathrm{rank}(A(k))$.

Also, if there exists a ν that satisfies (10), then there definitely exists a solution with atomic measure.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[2] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1277–1305, 2007.

[3] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[4] D. C. Dowson and A. Wragg, "Maximum entropy distributions having prescribed first and second moments," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 5, pp. 689–693, Sep. 1973.

[5] V. Poor, "Minimum distortion functional for one-dimensional quantisation," *Electron. Lett.*, vol. 16, no. 1, pp. 23–25, 1980.

[6] P. Ishwar and P. Moulin, "On the existence and characterization of the maxent distribution under general moment inequality constraints," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3322–3333, Sep. 2005.

[7] T. T. Georgiou, "Relative entropy and the multivariable multidimensional moment problem," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1052–1066, Mar. 2006.

[8] M. P. Friedlander and M. R. Gupta, "On minimizing distortion and relative entropy," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 238–245, Jan. 2006.

[9] A. Dukkipati, M. N. Murty, and S. Bhatnagar, "Nonextensive triangle equality and other properties of Tsallis relative entropy minimization," *Physica A: Statist. Mech. Its Appl.*, vol. 361, no. 1, pp. 124–138, 2006.

[10] G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *J. Mach. Learning Res.*, vol. 3, pp. 555–582, 2002.

[11] I. Csiszar, G. Tusnady, M. Ispany, E. Verdes, G. Michaletzky, and T. Rudas, "Divergence minimization under prior inequality constraints," in *Proc. IEEE Symp. Inf. Theory*, 2001, p. 21.

[12] I. Csiszar and F. Matus, "On minimization of multivariate entropy functionals," in *Proc. IEEE Symp. Inf. Theory*, 2009, pp. 96–100.

[13] A. Antos, L. Devroye, and L. Györfi, "Lower bounds for Bayes error estimation," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 21, no. 7, pp. 643–645, Jul. 1999.

[14] R. E. Curto and L. A. Fialkow, "Recursiveness, positivity, and truncated moment problems," *Houston J. Math.*, vol. 17, no. 4, pp. 603–635, 1991.

[15] D. Bertsimas and I. Popescu, "Optimal inequalities in probability theory: A convex optimization approach," *SIAM J. Optim.*, vol. 15, no. 3, pp. 780–804, 2005.

[16] A. W. Marshall and I. Olkin, "Multivariate Chebyshev inequalities," *The Ann. Math. Statist.*, vol. 31, no. 4, pp. 1001–1014, 1960.

Bela A. Frigyik received the Ph.D. in Mathematics from the University of Washington in 2006.

His undergraduate degree is from the Technical University of Budapest. From 2006–2009, he was a Research Assistant Professor of Mathematics at Purdue University.

Since 2009, he has been a Research Associate at the University of Washington.

Maya R. Gupta (SM'10) received the Ph.D. in Electrical Engineering from Stanford in 2003. Her undergraduate degrees were from Rice University. Since 2003, she has been an Associate Professor at the University of Washington.