Bounds on the Maximum Bayes Error Given Moments

Bela A. Frigyik and Maya R. Gupta

Abstract

We show how to compute lower bounds for the maximum possible Bayes error if the class-conditional distributions must satisfy moment constraints. Our approach makes use of Curto and Fialkow's solutions for the truncated moment problem. We also give an upper bound that follows from related work by Lanckreit et al., Bertsimas and Popescu, and Marshall and Olkin.

Index Terms

Bayes error, maximum entropy, moment constraint, truncated moments, quadratic discriminant analysis

I. Introduction

We assume that one is given the first n_i moments for the ith class-conditional distribution for $i = 1, \ldots, G$, and consider the question, "What is the maximum possible Bayes error given such moment constraints?" A common 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, known as *linear* or *quadratic discriminant analysis* (QDA) [1], [2]. While Gaussians are known to be maximize entropy given moment constraints [3], this research was motivated by our interest in knowing how robust the Gaussian assumption of QDA is in terms of maximizing the Bayes error.

There are a number of 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], where a key difference in their work is that they instead of assuming that moments are given, they take as given iid samples from the class-conditional distributions, and they lower bound how wrong on average any estimate of the Bayes error might be.

After reviewing the Bayes error, 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). First, a simple construction shows that if only the means are known, there exist distributions such that the Bayes error is arbitrarily close to $(G-1)/G$. Given the first two moments, we show how to compute a lower bound on the maximum Bayes error. In Section V, we show that the QDA assumption that the class-conditional distributions are Gaussian produces a much lower Bayes error than the worst-possible Bayes error given the moment constraints.

We also discuss the general case of constructing a lower bound given the first n moments. Then in Section IV, we show that the approach of Lanckreit et al. [10] can be used to provide an upper bound on the maximum Bayes error. We end with a discussion of the tightness of these bounds and some open questions.

II. Bayes Error

Let $\mathcal{X}$ be a feature 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 $x \in \mathcal{X}$ let

$$\hat{y}(x) \in \operatorname{argmax}_{i \in \mathcal{Y}} p(i|x), \tag{II.1}$$

Frigyik and Gupta are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 (e-mail: frigyik@uw.edu, gupta@ee.washington.edu). This work was supported by the United States PECASE Award.

where we use $\in$ in (II.1) because $\arg\max$ returns the set of (possibly multiple) classes that equally maximize $p(i|x)$.

For a given x the probability of misclassification is

$$P_e(x) = \sum_{i \neq \hat{y}(x)} p(i|x) = 1 - p(\hat{y}(x)|x).$$

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

$$\mathbb{E}[P_e] = 1 - \int p(\hat{y}(x)|x) d\nu(x) \tag{II.2}$$

Further suppose that the measure ν is defined on $\mathcal{X}$ such that it is either absolutely continuous w.r.t. Lebesgue measure, or is discrete and expressed

$$d\nu(x) = \sum_{j=1}^{\infty} \alpha_j \delta_{x_j}(x),$$

where δ_{x_j} is the Dirac measure with support $\{x_j\}$ and $\alpha_j > 0$ for all $j = 1, 2, \ldots$. Throughout this paper, we make the simplifying assumption that $p(i) = 1/G$ for all i. With this simplifying assumption we have the following (details in the Appendix). In the discrete case, (II.2) becomes

$$\mathbb{E}[P_e] = 1 - \frac{1}{G} \sum_{j=1}^{\infty} \max_{i \in \mathcal{Y}} \beta_j(i, x_j)$$

where $\beta_j(i, x_j) = G p(i|x_j) \alpha_j$; and in the absolutely continuous case,

$$\mathbb{E}[P_e] = 1 - \frac{1}{G} \int \max_{i \in \mathcal{Y}} p(x|i) dx. \tag{II.3}$$

III. Lower Bounds for Worst-Case Bayes Error

Our strategy to providing lower bounds is to constrain the G probability distributions 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. Then the Bayes error will be at least $\frac{G-1}{G}\epsilon$. The largest such ϵ for which this overlap constraint is feasible determines the best lower bound on the worst-case Bayes error we can provide. The maximum 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 given the first moment showing that no matter what the first moments are, the Bayes error can be arbitrarily bad. 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 III.1. *If the first moments $\gamma_{1,i}$ are given for every 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 a discrete or an absolutely continuous class-conditional distribution such that the Bayes error $\mathbb{E}[P_e] \geq \frac{G-1}{G}\epsilon$.*

Proof: The moment constraints hold if the ith class-conditional distribution is taken to be $\epsilon\delta_0 + (1-\epsilon)\delta_{z_i}$ 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 $\frac{G-1}{G}\epsilon$. For an absolutely continuous example, consider the uniform densities $p_l(s, i) = \frac{1}{2l}\mathbb{I}_{[\gamma_{1,i}-l, \gamma_{1,i}+l]}$ with $i = 1, 2, \ldots, G$. As $l \to \infty$ the Bayes error goes to $\frac{G-1}{G}$. ∎

Theorem III.1. *Suppose that the first and second moments are given for G class-conditional measures, i.e. for the ith class-conditional measure we are given $\{\gamma_{1,i}, \gamma_{2,i}\}$. A lower bound on the Bayes error is*

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

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

If $G = 2$ then

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

and the optimal Δ values are

$$\Delta_{1,2} = \frac{-(\gamma_{1,2}\gamma_{2,1} - \gamma_{1,1}^2\gamma_{1,2} + \gamma_{1,2}^2\gamma_{1,1} - \gamma_{2,2}\gamma_{1,1}) \pm \sqrt{(\gamma_{1,2}^2 - \gamma_{2,2})(\gamma_{1,1}^2 - \gamma_{2,1})(\gamma_{1,1} - \gamma_{1,2})^2}}{\gamma_{2,2} + \gamma_{1,1}^2 - \gamma_{2,1} - \gamma_{1,2}^2} \tag{III.1a}$$

if $\gamma_{2,2} + \gamma_{1,1}^2 - \gamma_{2,1} - \gamma_{1,2}^2 \neq 0$, and

$$\Delta = \frac{\gamma_{1,1}^2\gamma_{2,2} - \gamma_{1,2}^2\gamma_{2,1}}{2(\gamma_{1,1}\gamma_{2,2} + \gamma_{1,1}^2\gamma_{1,2} - \gamma_{1,2}\gamma_{2,1} - \gamma_{1,2}^2\gamma_{1,1})} \tag{III.1b}$$

otherwise.

Proof: Consider some $0 < \epsilon < 1$. A necessary condition for the Bayes error to be at least $(G/G-1)\epsilon$ is if each of the unknown measures shares 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 $\epsilon\delta_0 + \nu$ where $\nu(\{0\}) = 0$. If μ satisfy the original moment constraints then ν also satisfies them; this follows directly from the moment definition:

$$\int x^n d\mu(x) = 0^n\epsilon + \int x^n d\nu(x) = \int x^n d\nu(x).$$

Also $\nu(\mathcal{X}) = 1 - \epsilon$. Thus, we require a measure ν with a zeroth moment $\gamma_0 = 1 - \epsilon > 0$ and the original first and second moments γ_1, γ_2. Then, as described in Section VIII, 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 $\epsilon \leq 1 - \frac{\gamma_1^2}{\gamma_2}$. Moreover, the rank of matrix A and the rank of γ (for notation see Section VIII) 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 $\mathrm{rank}(\gamma) = 2$. We are not interested in the case that $\epsilon = 0$, therefore there is a measure ν with moments $\{1 - \epsilon, \gamma_1, \gamma_2\}$ iff $0 < \epsilon \leq 1 - \frac{\gamma_1^2}{\gamma_2}$.

Suppose we have G class-conditional measures and the corresponding moments constraints as given in the statement of this theorem. Let's denote the i*th* measure by $\epsilon_i\delta_0 + \nu_i$, where $\epsilon_i \leq 1 - \frac{\gamma_{1,i}^2}{\gamma_{2,i}}$. Then

$$\begin{aligned}
\max \mathbb{E}[P_e] &\geq 1 - \frac{1}{G}\left[\max_{i \in \mathcal{Y}}\{\epsilon_i\} + \sum_{i=1}^{G}(1 - \epsilon_i)\right] = \frac{1}{G}\left[\sum_{i=1}^{G}\epsilon_i - \max_{i \in \mathcal{Y}}\{\epsilon_i\}\right] \\
&\geq \frac{1}{G}\left[\sum_{i=1}^{G}\left(1 - \frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right) - \max_{i \in \mathcal{Y}}\left\{1 - \frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right\}\right] \\
&\geq \frac{G-1}{G}\min_{i \in \mathcal{Y}}\left\{1 - \frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right\},
\end{aligned} \tag{III.2a}$$

where the maximum 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 (III.2a) 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.

Shifting a distribution by Δ does change all of the moments (because they are not centered moments), specifically, if ν is a probability measure with finite moments $\gamma_0 = 1$, γ_1, ..., γ_n, and ν_Δ is the measure defined by $\nu_\Delta(D) = \nu(D + \Delta)$ for all ν-measurable sets D, then the n-th non-centered moment of the shifted measure ν_Δ is

$$\tilde{\gamma}_n = \int x^n d\nu_\Delta(x) = \int (x - \Delta)^n d\nu(x) = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \Delta^{n-k} \gamma_k,$$

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 (III.2a):

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

If $G = 2$ then

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

Since the functions $f_i(\Delta) = 1 - \frac{(\gamma_{1,i} - \Delta)^2}{\gamma_{2,i} + \Delta^2 - 2\Delta\gamma_{1,i}}$ have maximum at $\gamma_{1,i}$ and their derivative is strictly positive for $\Delta < \gamma_{1,i}$ and strictly negative for $\Delta > \gamma_{1,i}$ the potential maximum occurs at the point where the two functions are equal. This results in a quadratic equations if $\gamma_{2,2} + \gamma_{1,1}^2 - \gamma_{2,1} - \gamma_{1,2}^2 \neq 0$ with solutions (III.1a) or a linear one with solution (III.1b).

If we have more than one potential solution, after checking the derivative of the functions

$$f_i'(\Delta) = \frac{2(\Delta - \gamma_{1,i})(\gamma_{1,i}^2 - \gamma_2)}{(\gamma_{2,i} + \Delta^2 - 2\Delta\gamma_{1,i})^2}$$

at the potential Δ values, we can determine which one gives the maximum. ■

Corollary III.1. *Suppose 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 for any small $\delta > 0$ we can get a lower bound on the Bayes error that is not less than the lower bound we derived in Theorem III.1 minus $\frac{G-1}{G}\delta$, i.e.*

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

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 Section VIII imply that there is a measure with moments

$\{1-\epsilon, \gamma_1, \gamma_2, \gamma_3\}$ and

$$\max \mathbb{E}[P_e] \geq \frac{1}{G}\left[\sum_{i=1}^{G}\left(1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}-\delta\right)-\max_{i\in\mathcal{Y}}\left\{1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}-\delta\right\}\right]$$
$$\geq \frac{G-1}{G}\min_{i\in\mathcal{Y}}\left\{1-\frac{\gamma_{1,i}^2}{\gamma_{2,i}}\right\}-\frac{G-1}{G}\delta.$$

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

Lemma III.2. *Suppose that the first n moments are given for G class-conditional measures, i.e. for the ith class-conditional measure we are given $\{\gamma_{1,i}, \gamma_{2,i}, \ldots, \gamma_{n,i}\}$. 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:*

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

where the maximum on the left is taken over all the measures satisfying the moments constraints mentioned above.

Proof: The first part of the proof of Theorem III.1 is applicable in this case. To check whether there exist measures with the desired property we can apply the method described in [14]. ■

IV. Upper Bound for Worst-Case Bayes Error

Because the Bayes error is the best error over all decision boundaries, one approach to constructing an upper bound would be to restrict the set of considered decision boundaries to a set for which the worst-case error is easier to analyze. For example, 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 mean constraints. Here we show that this does in fact produce an upper bound on the maximum Bayes error.

Lanckreit et al. [10]'s analysis is a direct consequence of the following result by Bertsimas and Popescu [15] (which follows from a result by Marshall and Olkin [16]):

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

where the sup in (IV.1) is over all probability measures ν with domain $\mathbb{R}^d$, mean μ, and centered second moment Σ; and S is some convex set in the domain of ν. In words, (IV.1) says, "Over all possible probability distributions with the given mean and covariance, the largest probability that a random sample from ν will be in the convex set S is $1/(1+c(S))$."

Next, consider two fixed class-conditional measures ν_1, ν_2. As in Lanckreit et al. [10], consider the set of linear decision boundaries, which split the domain into two half-planes S_1, S_2. Then the error of a classifier corresponding to a linear decision boundary corresponding to the split (S_1, S_2) is

$$\frac{1}{2}\nu_1(S_2)+\frac{1}{2}\nu_2(S_1) \geq \text{ Bayes error}(\nu_1,\nu_2). \tag{IV.2}$$

As stated in (IV.2), the error with a linear decision boundary upper bounds the Bayes error for those two measures. To obtain a tighter upper bound, minimize the left-hand side over all linear decision boundaries:

$$\frac{1}{2}\inf_{S_1,S_2}(\nu_1(S_2)+\nu_2(S_1)) \geq \text{ Bayes error}(\nu_1,\nu_2). \tag{IV.3}$$

Now suppose ν_1 and ν_2 are unknown, but their first and second centered moments (μ_1, Σ_1) and (μ_2, Σ_2) are given. Then the worst-case error over all measures ν_1 and ν_2 for a fixed linear decision boundary S_1, S_2 is

$$\sup_{\nu_1|\mu_1,\Sigma_1}\nu_1(S_2)+\sup_{\nu_2|\mu_2,\Sigma_2}\nu_2(S_1).$$

We upper bound the worst-case Bayes error:

$$\begin{aligned}\text{worst-case Bayes error } &\leq \frac{1}{2} \sup_{\nu_1|\mu_1,\Sigma_1} \sup_{\nu_2|\mu_2,\Sigma_2} \inf_{S_1,S_2} (\nu_1(S_2) + \nu_2(S_1)) \\ &\leq \frac{1}{2} \inf_{S_1,S_2} \left(\sup_{\nu_1|\mu_1,\Sigma_1} \nu_1(S_2) + \sup_{\nu_2|\mu_2,\Sigma_2} \nu_2(S_1) \right). \end{aligned} \tag{IV.4}$$

Consider the one-dimensional case, in which all decision boundaries are merely points $s \in \mathbb{R}$, and hence all decision boundaries are linear. Also, 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 (IV.1) simplify (for details, see Appendix A of [10]), so that (IV.4) becomes

$$\text{maximum Bayes error } \leq \frac{1}{2} \inf_{s \in [0,\mu_2]} \left(\frac{1}{1 + \frac{s^2}{\sigma_1^2}} + \frac{1}{1 + \frac{(\mu_2 - s)^2}{\sigma_2^2}} \right). \tag{IV.5}$$

V. Comparison to Error With Gaussians

We illustrate the bounds described in this paper for the common case of two given moments. 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 10 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 infinum in (IV.5) occurs at $s = \mu_2/2$ and the upper bound is $4/(4 + \mu_2^2)$. For the second case with different variances we compute (IV.5) numerically.

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 Open Questions

Overall, we note that lower bounds for the worst-case Bayes error can be constructed by constraining the distributions. For example, constraining the distributions to be Gaussians produces a weak lower bound, and our tighter lower bound was constructed 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 of $G/G - 1$. 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 we do not control all the possible overlap between the class-conditional distributions, we believe it should be possible to construct tighter lower bounds for the $n \geq 2$ moment cases.

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 directly to an upper bound on the maximum Bayes error. However, the inequality introduced in (IV.4) when we switched the $\inf$ and $\sup$ may make this upper bound loose. We believe this upper bound will be tightest for $d = 1$ dimensions, because then the set of linear decision boundaries is the set of all decision boundaries. However, for $d > 1$ the class of linear decision boundaries only approximates the set of all possible decision boundaries, so (IV.3) may be quite loose.

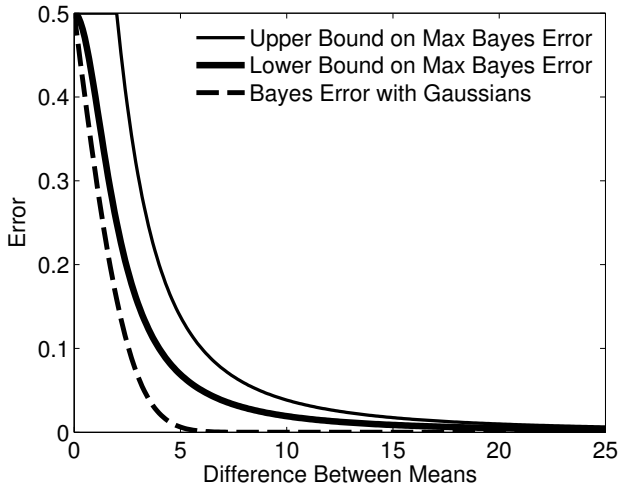
In practice, a moment constraint is often created by estimating the moment from samples drawn iid from the distribution. In that case, the moment constraint need not be treated as a hard constraint as we have done here. Rather, the observed samples can imply a probability distribution over the moments, 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.

Appendix

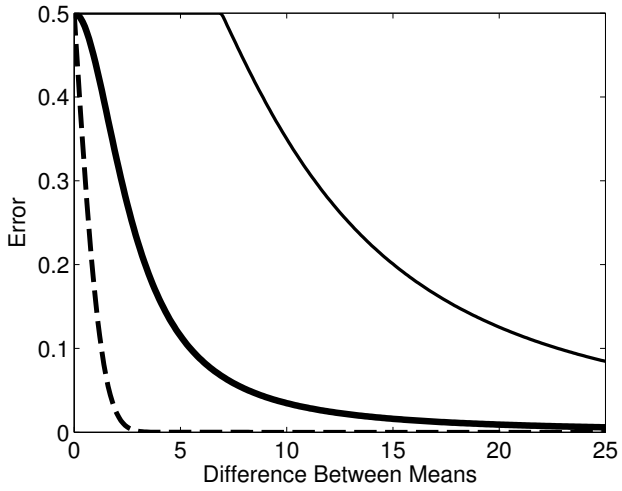
VII. Simplifying the Bayes Error

We detail how (II.2) becomes (II.3); the discrete argument is analogous. Let A_i be the set of all the x values such that $\hat{y}(x) = i$, i.e. $A_i = \{x : \hat{y}(x) = i\}$. Then

$$\hat{y}(x) = \sum_{i=1}^{G} \mathbb{I}_{A_i}(x) i,$$



$\sigma_2^2 = 1$



$\sigma_2^2 = 5$

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

and

$$\mathbb{E}[P_e] = 1 - \sum_{i=1}^{G} \int_{A_i} p(i|x) d\nu(x).$$

In this case there is a density associated with the measure and $d\nu(x) = p(x)dx$ therefore

$$p(i|x)d\nu(x) = p(i|x)p(x)dx$$

is the measure on $\mathcal{X} \times \mathcal{Y}$ with density $p(x, i) = p(i|x)p(x)$. Bayes' theorem says that $p(i|x)p(x) = p(x|i)p(i)$ where $p(i) = \int p(i|x)p(x)dx$. This means that

$$\mathbb{E}[P_e] = 1 - \sum_{i=1}^{G} \int_{A_i} p(i|x) d\nu(x) = 1 - \sum_{i=1}^{G} p(i) \int_{A_i} p(x|i) dx.$$

Assume that $p(i) = p(j) = 1/G$ for all $i, j \in \mathcal{Y}$ then for all $x \in A_i$ and for all $k \in \mathcal{Y}$

$$p(x|i) = Gp(x)p(i|x) \geq Gp(x)p(k|x) = p(x|k),$$

hence $p(x|i) = \max_{j \in \mathcal{Y}} p(x|j)$ and

$$\mathbb{E}[P_e] = 1 - \sum_{i=1}^{G} p(i) \int_{A_i} p(x|i) dx = 1 - \frac{1}{G} \int \max_{i \in \mathcal{Y}} p(x|i) dx.$$

VIII. Existence and Uniqueness 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

$$\gamma_i = \int s^i d\nu(s)? \tag{VIII.1}$$

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 $n = 3$, $k = 1$ and

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

Let $\mathbf{v}_j$ be the $j+1th$ column of $A(k)$ for $0 \leq j \leq k$. 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 the linear combination of $\{\mathbf{v}_0, \ldots, \mathbf{v}_{r-1}\}$.

Then whether there exists a ν that satisfies (VIII.1) 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 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 (VIII.1), 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: John Wiley and Sons, 1991.
[4] D. C. Dowson and A. Wragg, "Maximum entropy distributions having prescribed first and second moments," *IEEE Trans. on Information Theory*, vol. 19, no. 5, pp. 689–693, 1973.
[5] V. Poor, "Minimum distortion functional for one-dimensional quantisation," *Electronics Letters*, 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. on Information Theory*, vol. 51, no. 9, pp. 3322–3333, 2005.
[7] T. T. Georgiou, "Relative entropy and the multivariable multidimensional moment problem," *IEEE Trans. on Information Theory*, vol. 52, no. 3, pp. 1052–1066, 2006.
[8] M. P. Friedlander and M. R. Gupta, "On minimizing distortion and relative entropy," *IEEE Trans. on Information Theory*, vol. 52, no. 1, pp. 238–245, 2006.
[9] A. Dukkipati, M. N. Murty, and S. Bhatnagar, "Nonextensive triangle equality and other properties of Tsallis relative entropy minimization," *Physica A: Statistical Mechanics and Its Applications*, vol. 361, no. 1, pp. 124–138, 2006.
[10] G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *Journal Machine Learning Research*, 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," *IEEE Symp. Information Theory*, 2001.
[12] I. Csiszar and F. Matus, "On minimization of multivariate entropy functionals," *IEEE Symp. Information Theory*, 2008.
[13] A. Antos, L. Devroye, and L. Györfi, "Lower bounds for Bayes error estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 7, pp. 643–645, 1999.
[14] R. E. Curto and L. A. Fialkow, "Recursiveness, positivity, and truncated moment problems," *Houston Journal of Mathematics*, vol. 17, no. 4, pp. 603–635, 1991.
[15] D. Bertsimas and I. Popescu, "Optimal Inequalities in Probability Theory: A Convex Optimization Approach," *SIAM Journal on Optimization*, vol. 15, no. 3, pp. 780–804, 2005.
[16] A. W. Marshall and I. Olkin, "Multivariate Chebyshev inequalities," *The Annals of Mathematical Statistics*, vol. 31, no. 4, pp. 1001–1014, 1960.