The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter

Vittorio Castelli, *Member, IEEE*, and Thomas M. Cover, *Fellow, IEEE*

***Abstract*— We observe a training set Q composed of l labeled samples $\{(X_1,\theta_1),\cdots,(X_l,\theta_l)\}$ and u unlabeled samples $\{X'_1,\cdots,X'_u\}$. The labels θ_i are independent random variables satisfying $\Pr\{\theta_i=1\}=\eta$, $\Pr\{\theta_i=2\}=1-\eta$. The labeled observations X_i are independently distributed with conditional density $f_{\theta_i}(\cdot)$ given θ_i. Let (X_0,θ_0) be a new sample, independently distributed as the samples in the training set. We observe X_0 and we wish to infer the classification θ_0. In this paper we first assume that the distributions $f_1(\cdot)$ and $f_2(\cdot)$ are given and that the mixing parameter η is unknown. We show that the relative value of labeled and unlabeled samples in reducing the risk of optimal classifiers is the ratio of the Fisher informations they carry about the parameter η. We then assume that two densities $g_1(\cdot)$ and $g_2(\cdot)$ are given, but we do not know whether $g_1(\cdot)=f_1(\cdot)$ and $g_2(\cdot)=f_2(\cdot)$ or if the opposite holds, nor do we know η. Thus the learning problem consists of both estimating the optimum partition of the observation space and assigning the classifications to the decision regions. Here, we show that labeled samples are necessary to construct a classification rule and that they are exponentially more valuable than unlabeled samples.**

***Index Terms*— Pattern recognition, supervised learning, unsupervised learning, labeled and unlabeled samples, Bayesian method, Laplace's integral, asymptotic theory.**

I. Introduction

We ADDRESS the problem of the optimal use of a training set Q composed of l labeled samples $\{(X_1,\theta_1),\cdots,(X_l,\theta_l)\}$ and u unlabeled samples $\{X'_1,\cdots,X'_u\}$ in the construction of a classifier that discriminates between two classes of observations, called Class 1 and Class 2, and we propose a solution under two particular sets of hypotheses. We denote the prior probability of observing a sample of Class 1 by η, and the densities of the observations of Class 1 and Class 2 by $f_1(\cdot)$ and $f_2(\cdot)$, respectively. We analyze the behavior of $R(l,u)$, the probability of classification error of an optimal procedure based on a training set that contains l labeled samples and u unlabeled samples.

We have analyzed this problem in previous work [2] under the assumption that the density of the unlabeled samples is identifiable, and we have considered training sets composed of an infinite number of unlabeled samples and a finite number of labeled samples. Optimum classification rules exist in such a framework. Using the notation $R(l,u)$ to denote the risk of such optimum rules, it is shown in [2] that

$$R(0,u)=R(0,\infty)=1/2, \qquad \text{for all } u$$

and thus labeled samples are necessary to construct a classifier. It is also shown in [2] that when the training set contains an infinite number of unlabeled samples and one labeled sample, the probability of error of the optimum classifier is given by

$$R(1,\infty)=2R^*(1-R^*)$$

where

$$R^*=\int \min\{\eta f_1(x),\overline{\eta}f_2(x)\}\,dx$$

denotes the Bayes risk. Finally, it is also shown in [2] that additional labeled samples make the probability of error converge exponentially fast to the Bayes risk in the sense that

$$R(l,\infty)-R^*=\exp\{-l\,D+o(l)\}.$$

We have proved that the exponent D is equal to

$$-\log\left\{2\sqrt{\eta\overline{\eta}}\int\sqrt{f_1(x)f_2(x)}\,dx\right\}$$

where $\overline{\eta}$ is a shorthand notation for $1-\eta$.

In this paper, we first assume that the densities $f_1(\cdot)$ of the samples of Class 1 and $f_2(\cdot)$ of the samples of Class 2 are given and that the probability η of observing a sample of Class 1 is unknown. Using a result by O'Neill [21], we show that the risk $R(l,u)$ of a Bayesian classifier with positive and continuous prior $h(\eta)$ over $[0,1]$ satisfies the asymptotic relation

$$\Delta R(l,u)\triangleq R(l,u)-R^*=\frac{c}{lI_l+uI_u}(1+o(1))$$

where c is a function of η, $f_1(\cdot)$, and $f_2(\cdot)$ but not of the prior $h(\cdot)$. We conclude that the first-order term in the expansion of $\Delta R(l,u)$ is the same for all nondegenerate Bayesian classifiers and that the relative value of labeled and unlabeled samples in reducing $R(l,u)$ is equal to the ratio of their Fisher informations. Moreover, labeled samples are not necessary in

Manuscript received December 16, 1994; revised October 6, 1995. This work was partially supported by NSF under Grant NCR-8914538-02, by ARPA under Contract J-FBI-91-218, and by JSEP under Contract DAAHO4-94-G-0058, and was done when both authors were with Stanford University. The material in this paper was presented in at the International Symposium on Information Theory, Trondheim, Norway, 1994.

V. Castelli is with IBM T. J. Watson Research Center, Yorktown Hights, NY 10598 USA.

T. M. Cover is jointly in the Department of Electrical Engineering and Statistics, Stanford University, Stanford, California 94305 USA.

Publisher Item Identifier S 0018-9448(96)06370-5.

0018–9448/96$05.00 © 1996 IEEE

this framework and one can construct a decision rule based solely on unlabeled observations, the risk of which converges to R^* with rate u^{-1}.

The main result of this paper is based on the assumption that two densities $g_1(\cdot)$ and $g_2(\cdot)$ are given, and that we do not know whether $g_1(\cdot) = f_1(\cdot)$ or $g_1(\cdot) = f_2(\cdot)$. As before, the mixing parameter η is unknown. We show that labeled samples are now necessary to construct a classifier, and we prove that, when $l^{3+\epsilon}u^{-1} \to 0$ as $l+u \to \infty$, the risk $R(l,u)$ satisfies

$$\Delta R(l,u) = R(l,u) - R^* = O(u^{-1}) + \exp\{-lD + o(l)\}$$

where

$$D = -\log\left\{2\sqrt{\eta\overline{\eta}} \int \sqrt{f_1(x)f_2(x)}\, dx\right\}.$$

This implies that, if $u\exp\{-D\ l\} \to 0$ and $l = o(u)$, the additional risk $\Delta R(l,u)$ is essentially determined by the number of unlabeled samples in the training set, while if the number of unlabeled samples u grows faster than $\exp\{D\ l\}$, then $\Delta R(l,u)$ is essentially determined by the number of labeled samples.

These results should extend to the more general framework where the mixing parameter η and the the densities $f_1(\cdot)$ of the samples of Class 1 and $f_2(\cdot)$ of the samples of Class 2 are unknown, $f_1(\cdot)$ and $f_2(\cdot)$ belong to a regular parametric family $\mathcal{F}$, and the class of mixtures $\mathcal{G}$ of two elements of $\mathcal{F}$ is identifiable and satisfies the conditions for Laplace regularity [13]. We can show [1] that the main result of the present work still holds, namely, that labeled samples are exponentially more valuable than unlabeled samples in reducing the probability of error of nondegenerate Bayesian classifiers.

The general problem of learning with both labeled and unlabeled observations is of practical relevance. In fact, the construction of a training set is often performed by collecting observations first and by labeling them afterwards, and in many instances the labeling process is harder or more expensive than the sampling step required to obtain the observations. Sometimes it is impossible to label most samples, and in some cases it is desirable to keep to a minimum the number of labeled samples, as for example when labeling involves the death of the patient. Therefore, situations in which both labeled and unlabeled samples are available arise naturally, and the investigation of the simultaneous use of both kinds of observations in learning leads immediately to questions of the relative value of labeled and unlabeled samples.

Among the practical learning schemes that use training sets composed of both labeled and unlabeled samples there are neural network classifiers [20], [23], [10], [11], [5], 28] and parametric methods in discriminant analysis [22], [17], [18], [26]. The problem of supervised learning in a parametric setting has been widely studied, the most commonly encountered approaches being maximum-likelihood parameter estimation and Bayesian parameter estimation [19, ch. 2.1], [4, ch. V.5], [9, ch. 5], [27, ch. 7]. Unsupervised learning is also often addressed in a parametric setting [8, ch. 6], [4, ch. V]. Many studies exist that analyze the behavior of the risk of classifiers in a supervised learning framework both for small training set size [12], [24] and for large training set size [7], [27, ch. 9].

II. Framework and Preliminaries

Denote the *labeled* samples by pairs (X_i, θ_i), where the class labels $\{\theta_i\}$ are independent and identically distributed with

$$\Pr\{\theta_i = 1\} = \eta \qquad \Pr\{\theta_i = 2\} = 1 - \eta \triangleq \overline{\eta}$$

and the observations $\{X_i\}$ are independent m-dimensional feature vectors with conditional densities $f_{\theta_i}(\cdot)$ given θ_i. Thus the joint density of the labeled pairs is

$$f_{X,\theta}(x,\theta) = [\eta f_1(x)]^{1(\theta=1)}[\overline{\eta} f_2(x)]^{1(\theta=2)} = \eta_\theta f_\theta(x) \quad (1)$$

where the indicator function $1(\theta = k)$ is equal to 1 if $\theta = k$, and to 0 otherwise, and where $\eta_1 = \eta, \eta_2 = \overline{\eta}$. It is understood to be a density with respect to the usual product of Lebesgue measure for x in $\mathbb{R}^m$ and counting measure for θ in $\{1, 2\}$.

When only the observation part X is known, we say that a sample is *unlabeled.* To indicate an unlabeled observation we use a prime superscript. The unlabeled observations X'_j appear to be distributed according to a mixture distribution with density

$$f_{X'}(\cdot) = \eta f_1(\cdot) + \overline{\eta} f_2(\cdot) \quad (2)$$

which is the marginal of X corresponding to the joint (1). To indicate dependence on η we also write these densities as $f(x,\theta|\eta)$ and $f(x'|\eta)$, respectively.

Observe a training set Q composed of l labeled samples $\{(X_1,\theta_1),\cdots,(X_l,\theta_l)\}$ and u unlabeled samples $\{X'_1,\cdots,X'_u\}$. Let (X_0,θ_0) be a new sample independently distributed as the samples in the training set. We want to infer the classification θ_0 from the training set Q and from X_0, and we wish to minimize the probability of error $\Pr\{\hat{\theta}_0(X_0,Q) \neq \theta_0\}$ among the class of measurable classification rules

$$\{\hat{\theta}_0(X_0,Q)\colon \mathbb{R}^m \times \{Q\} \to \{1,2\}\}$$

where $\{Q\}$ refers to the set of possible training sets with $(X_i,\theta_i) \in \mathbb{R}^m \times \{1,2\}$ for $i = 1,2,\cdots,l$ and X'_j in $\mathbb{R}^m$ for $j = 1,2,\cdots,u$.

When $\eta, f_1(\cdot)$ and $f_2(\cdot)$ are known, the Bayes decision rule is optimum for classifying X_0. The Bayes decision rule is the likelihood ratio test

$$\begin{aligned} &\text{decide } \hat{\theta}_0(X_0) = 1 \quad \text{if } \frac{f_1(X_0)}{f_2(X_0)} > \frac{\overline{\eta}}{\eta} \\ &\text{decide } \hat{\theta}_0(X_0) = 2 \quad \text{if } \frac{f_1(X_0)}{f_2(X_0)} < \frac{\overline{\eta}}{\eta}. \end{aligned} \quad (3)$$

The corresponding probability of error, the Bayes risk R^*, is given by

$$R^* = \int_{\mathbb{R}^m} \min\{\eta f_1(x), \overline{\eta} f_2(x)\}\, dx. \quad (4)$$

and is a lower bound to the probability of error of any classification rule.

When the training set is finite, the two densities $f_1(\cdot)$ and $f_2(\cdot)$ are known, and the mixing parameter η is unknown, there is no uniformly optimum rule that discriminates between the alternatives $\theta_0 = 1$ and $\theta_0 = 2$. A general discussion of this matter can be found, for instance, in Lehmann [16, ch. 1.4].

Here we are interested in the behavior of the risk of consistent classifiers, those with risk converging to R^* for all values of the mixing parameter η. Among the possible definitions of optimality (see, for example, the review in SenGupta [25]), we choose to restrict the attention to admissible procedures. A test T is admissible if no other test exists with risk smaller than or equal to the risk of T for every η and strictly smaller on a nonempty set. Among the class of consistent and admissible classification rules, we consider Bayes tests with smooth prior densities $h(\eta)$ over $[0,1]$ and calculate the rate of convergence of the corresponding risk to R^*. This choice gives us a sufficiently rich family of classifiers, which are optimal in the sense described in Lehmann [16, ch. 16]. Note that in a Bayesian framework, the mixing parameter η is a random variable. For notational purposes we shall use η to indicate both the random variable and the dummy variable in integration and differentiation, and we shall use η_0 to indicate the actual value of the mixing parameter, i.e., the realization of the random variable η.

Then, in the current framework $R(l,u)$ will denote the risk of a Bayesian classifier based on a training set Q composed of l labeled samples and u unlabeled samples, and $\Delta R(l,u)$ will be a shorthand for $(R(l,u) - R^*)$. Clearly, $\Delta R(l,u)$ depends on the choice of the prior $h(\cdot)$, and arguably $\Delta R_h(l,u)$ would be a more appropriate notation. Nevertheless, if the prior $h(\cdot)$ satisfies the regularity conditions specified in the theorems, the first-order terms in the asymptotic expansion of $\Delta R(l,u)$ as a function of l and u do not depend on the prior, thus justifying the notation. In addition, there is yet another advantage in the choice of a Bayesian framework for the analysis, that provides a rich family of optimal and admissible classifiers for which the large sample size behavior of the additional risk $\Delta R(l,u)$ independent of the choice of the prior.

A Bayesian solution to a classification problem is by definition a rule that minimizes the conditional probability of error given the training set Q and the new observation X_0 (posterior probability of error) among the class of measurable functions

$$\{\hat{\theta}(X_0,Q)\colon \mathbb{R}^m \times \{Q\} \to \{1,2\}\}.$$

The posterior probability of error can be written as

$$\begin{aligned} P_e(X_0,Q) &\triangleq \Pr\{\hat{\theta}(X_0,Q) \neq \theta_0 | X_0, Q\} \\ &= \Pr\{\hat{\theta}(X_0,Q) = 1\} \Pr\{\theta_0 = 2 | X_0, Q\} \\ &\quad + \Pr\{\hat{\theta}(X_0,Q) = 2\} \Pr\{\theta_0 = 1 | X_0, Q\}. \quad (5) \end{aligned}$$

Note that the overall probability of error P_e for a given rule $\hat{\theta}$ is given by

$$P_e = \int P_e(X_0,Q) f(X_0,Q)\, dX_0\, dQ.$$

where $f(X_0,Q) = f(X_0)f(Q)$ with

$$f(Q) = \prod_{i=1}^{l} f(X_i,\theta_i) \cdot \prod_{j=1}^{u} f(X'_j)$$

and the integration is understood in the natural way to be integration with respect to a product of Lebesgue and counting measures for (X_i,θ_i) and Lebesgue measure for X'_i and X_0.

The classifier $\hat{\theta}_0(X_0,Q)$ minimizing (5) is then characterized by

$$\begin{aligned} &\text{decide } \hat{\theta}_0(X_0,Q) = 1 \\ &\qquad \text{if } \Pr\{\theta_0 = 1|X_0,Q\} > \Pr\{\theta_0 = 2|X_0,Q\} \\ &\text{decide } \hat{\theta}_0(X_0,Q) = 2 \\ &\qquad \text{if } \Pr\{\theta_0 = 1|X_0,Q\} < \Pr\{\theta_0 = 2|X_0,Q\} \end{aligned}$$

and is unique except on the set

$$\{\Pr\{\theta_0 = 1|X_0,Q\} = \Pr\{\theta_0 = 2|X_0,Q\}\}.$$

Consequently, a Bayesian solution is a ratio test of the form

$$\begin{aligned} &\text{decide } \hat{\theta}_0(X_0,Q) = 1 \quad \text{if } \frac{\Pr\{\theta_0 = 1|X_0,Q\}}{\Pr\{\theta_0 = 2|X_0,Q\}} > 1 \\ &\text{decide } \hat{\theta}_0(X_0,Q) = 2 \quad \text{if } \frac{\Pr\{\theta_0 = 1|X_0,Q\}}{\Pr\{\theta_0 = 2|X_0,Q\}} < 1. \quad (6) \end{aligned}$$

The following notation is used throughout this paper. Vectors are meant to be column vectors. We use the same notation to indicate scalars and vectors as it is always clear from the context whether quantities are scalar- or vector-valued. Distinct points in a d-dimensional Euclidean space will be identified by different subscripts. To denote the ith component of a vector x we use parentheses around the subscript. Thus for instance, x_1 and x_2 are distinct points in an appropriate Euclidean space $\mathbb{R}^m$, and $\partial/\partial x_{(i)}$ is the partial derivative with respect to the ith component of x. ∇ denotes the gradient and $\|\cdot\|$ the Euclidean norm. The overline sign is used in two distinct contexts. If a random variable θ takes value in the set $\{1,2\}$, then $\overline{\theta} \triangleq 1$ if $\theta = 2$ and $\overline{\theta} \triangleq 2$ if $\theta = 1$. If η takes value in the interval $[0,1]$, then $\overline{\eta} \triangleq 1 - \eta$.

III. Unknown Mixing Parameter η

The first problem we analyze is the construction of a classifier using a training set Q composed of l labeled samples $\{(X_1,\theta_1),\cdots,(X_l,\theta_l)\}$ and u unlabeled samples $\{X'_1,\cdots,X'_u\}$ under the assumption that the densities $f_1(\cdot)$ of the observations of Class 1 and $f_2(\cdot)$ of the observations of Class 2 are known, and that the actual value of the mixing parameter η_0 is unknown. We consider Bayesian classifiers with respect to a prior density $h(\cdot)$ over η.

Theorem 1, below, states that Bayesian solutions with respect to smooth priors are asymptotically equivalent, in the sense that for large sample size the additional risk $\Delta R(l,u)$ is, to the first order, independent of the choice of the prior density $h(\cdot)$. In addition, the theorem states that labeled and unlabeled samples play a very similar role in reducing the probability of classification error and that their relative value is equal to the ratio of the respective Fisher informations. The proof of Theorem 1 is a specialization of a result of O'Neill [21, Theorem 1], the details of which we shall omit. A different proof can be found in [3].

Theorem 1: Assume that the observations are real-valued random variables. Let the densities $f_{1(\cdot)}$ and $f_2(\cdot)$ be three times continuously differentiable with bounded first and second derivatives. Assume that, with the exception of a finite number of values of the mixing parameter η, there is a finite number of solutions $\{x_i^\eta\}$ of the equation $\eta f_1(\cdot) - \overline{\eta} f_2(\cdot) = 0$, and that the derivative of $[\eta f_1(x) - \overline{\eta} f_2(x)]$ with respect to x evaluated at each x_i^η is different from zero. Let Ξ be the exceptional subset of $[0, 1]$ where the previous conditions are not satisfied. Assume that the actual value η_0 of the mixing parameter does not belong to Ξ, and that $0 < \eta_0 < 1$. Consider any Bayes test defined as in (9) and characterized by a prior distribution $h(\eta)$ that is continuous and positive at all $\eta \in [0, 1]$. The corresponding probability of error satisfies

$$R(l, u) - R^* \sim \frac{c}{l\, I_l(\eta_0) + u\, I_u(\eta_0)} \tag{7}$$

where

$$I_l(\eta_0) = \frac{1}{\eta_0 \overline{\eta} z}$$

and

$$I_u(\eta_0) = \int \frac{[f_1(x) - f_2(x)]^2}{\eta_0 f_1(x) + \overline{\eta} z f_2(x)}\, dx$$

are, respectively, the Fisher informations of the labeled and unlabeled samples, and c is the constant

$$c = \frac{1}{2} \sum_i \frac{[f_1(x_i^0) + f_2(x_i^0)]^2}{|\eta_0 f_1'(x_i^0) - \overline{\eta} z f_2'(x_i^0)|} \tag{8}$$

where the $\{x_i^0\}$ are the solutions of $[\eta_0 f_1(x) - \overline{\eta} z f_2(x)] = 0$.

The notation $\sim$ should be read "is asymptotically equivalent to" and means that the ratio of the right-hand side to the left-hand side converges to one. Note that c depends only on $\eta_0, f_1(\cdot)$ and $f_2(\cdot)$, but not on the prior $h(\cdot)$. Also, the only smoothness conditions on the prior $h(\cdot)$ required by Theorem 1 are $h(\cdot) \in C^0([0, 1])$ and $h(\eta) > 0$ for all $\eta \in [0, 1]$. Finally, note that if the observations X_i take values in $\mathbb{R}^m$, the constant c is replaced by

$$B = \frac{1}{2} \int_D \frac{(f_1(x) + f_2(x))^2}{\|\nabla[\eta_0 f_1(x) - \overline{\eta}_0 f_2(x)]\|}\, d\mu_D$$

where D is the surface of the equation $\eta_0 f_1(x) = \overline{\eta}_0 f_2(x)$ and μ_D is the Lebesgue measure on D.

Proof: We reduce the Bayesian classification problem to a Bayesian estimation problem with respect to the squared error loss function, by means of the following lemma.

Lemma 1: Bayesian solutions have the form

$$\begin{aligned} &\text{decide } \hat{\theta}_0 = 1 \quad \text{if } \frac{f_1(X_0)}{f_2(X_0)} > \frac{E_h[\overline{\eta}|Q]}{E_h[\eta|Q]} \\ &\text{decide } \hat{\theta}_0 = 2 \quad \text{if } \frac{f_1(X_0)}{f_2(X_0)} < \frac{E_h[\overline{\eta}|Q]}{E_h[\eta|Q]} \end{aligned} \tag{9}$$

where $h(\cdot)$ is the prior density on η, where the posterior expectation of η is

$$E_h[\eta|Q] = \int \eta f_h(\eta|Q)\, d\eta = \frac{\displaystyle\int \eta f(Q|\eta) h(\eta)\, d\eta}{f(Q)}$$

and the terms $f(Q|\eta)$ and $f(Q)$ are given by

$$\begin{aligned} f(Q|\eta) &= \prod_{i=1}^{l} \eta_{\theta_i} f_{\theta_i}(X_i) \prod_{j=1}^{u} [\eta f_1(X_j') + \overline{\eta} f_2(X_j')] \\ f(Q) &= E_h[f(Q|\eta)] \end{aligned}$$

where $\eta_1 = \eta$ and $\eta_2 = \overline{\eta}$.

The proof is given in Appendix A1. □

Theorem 1 has now been reduced to a classical framework. We refer the interested reader to O'Neill [21], where the behavior of parametric classifiers under regularity conditions is analyzed. Here, it suffices to remark that, thanks to Lemma 1, Theorem 1 can be derived as a specialization of O'Neill's Theorem 1 [21]. Moreover, the hypotheses of his theorem are satisfied by our current assumptions. □

Equation (7) can be interpreted as a Taylor series expansion of the risk around the true value of the parameter [3]. In fact, if $R(\hat{\eta})$ indicates the error rate of the likelihood ratio classifiers based on $\hat{\eta}$, instead of the true value of the mixing parameter η_0, then

$$\begin{aligned} R(\hat{\eta}) = {} & R(\eta_0) + \frac{d}{d\eta} R(\eta)|_{\eta_0} (\hat{\eta} - \eta_0) \\ & + \frac{1}{2} \frac{d^2}{d\eta^2} R(\eta)|_{\eta_0} (\hat{\eta} - \eta_0)^2 + O((\hat{\eta} - \eta_0)^3) \end{aligned}$$

and

$$R(l, u) = E[R(\hat{\eta})] = R^* + \frac{1}{2} \frac{d^2}{d\eta^2} R(\eta)|_{\eta_0} E[(\hat{\eta} - \eta_0)^2](1 + o(1)) \tag{10}$$

where

$$c = 2 \frac{d^2}{d\hat{\eta}^2} R(\hat{\eta})|_{\eta_0}$$

and

$$E[(\hat{\eta} - \eta_0)^2] \sim [l\, I_l(\eta_0) + u\, I_u(\eta_0)]^{-1}.$$

Here the convergence of the moments can be checked for the Bayesian estimator of η by modifying slightly the proof of a theorem [15, Theorem 6.7.1] of Lehmann. It is easy to show that the Fisher information for the labeled and the unlabeled samples are, respectively

$$I_l(\eta) = \frac{1}{\eta_0 \overline{\eta}_0} \qquad I_u(\eta) = \int \frac{[f_1(x) - f_2(x)]^2}{\eta_0 f_1(x) + \overline{\eta}_0 f_2(x)}\, dx.$$

From the inequalities

$$\begin{aligned} \frac{[f_1(x) - f_2(x)]^2}{\eta_0 f_1(x) + \overline{\eta}_0 f_2(x)} &\le \frac{f_1^2(x)}{\eta_0 f_1(x) + \overline{\eta}_0 f_2(x)} \\ &\quad + \frac{f_2^2(x)}{\eta_0 f_1(x) + \overline{\eta}_0 f_2(x)} \\ &\le \frac{f_1(x)}{\eta_0} + \frac{f_2(x)}{\overline{\eta}_0} \end{aligned}$$

it follows that $I_l(\eta_0) \geq I_u(\eta_0)$, and that equality holds if and only if the intersection of the support sets of $f_1(\cdot)$ and $f_2(\cdot)$ has measure zero. Thus labeled samples are strictly more valuable than unlabeled samples, unless the support sets of the underlying densities do not intersect.

A. An Example: Known $f_1(\cdot), f_2(\cdot)$, Unknown η

To illustrate the result, consider a simple example. Let $f_1(x) = 2x, f_2(x) = 2(1-x)$ for $x \in [0,1]$, $f_1(x) = f_2(x) = 0$ otherwise. Assume that the actual value of the mixing parameter is $\eta_0 = 1/2$. Then $\eta_0 f_1(x) + \overline{\eta}_0 f_2(x) = 1$ for $x \in [0,1]$. Here $\eta_0 f_1(\cdot)$ and $\overline{\eta}_0 f_2(\cdot)$ cross at a single point $x^0 = 1/2$. The constant c is then given by

$$\begin{aligned} c &= \frac{1}{2} \sum_{\{x_i^0\}} \frac{(f_1(x_i^0) + f_2(x_i^0))^2}{|\eta_0 f_1'(x_i^0) - \overline{\eta}_0 f_2'(x_i^0)|} \\ &= \frac{1}{2} \frac{(f_1(\frac{1}{2}) + f_2(\frac{1}{2}))^2}{|\frac{1}{2} f_1'(\frac{1}{2}) - \frac{1}{2} f_2'(\frac{1}{2})|} = \frac{1}{2} \frac{(1+1)^2}{|\frac{1}{2}2 + \frac{1}{2}2|} = 1 \end{aligned}$$

and the Fisher informations for labeled and unlabeled observations are, respectively

$$I_l(\eta_0) = \frac{1}{\eta_0 \overline{\eta}_0} = 4 \quad I_u(\eta_0) = \int_0^1 (2x - 2 + 2x)^2 \, dx = \frac{4}{3}.$$

Thus

$$R(l,u) - R^* \sim \frac{1}{4l + \frac{4}{3}u}$$

and consequently labeled samples are three times more valuable than unlabeled samples in reducing the probability of error. Adding one labeled sample to the training set results in the same reduction of the risk as adding three unlabeled samples. □

If the distributions of the two classes of samples are given, the only unknown being the mixing parameter η, the labeled and the unlabeled samples play a similar role in reducing the probability of error, in that both kinds of samples reduce the uncertainty in recovering the boundaries of the decision regions. A classification rule can be constructed using either labeled observations or unlabeled observations or both.

To obtain the same reduction in the probability of error that results from adding one labeled sample to the training set we must add I_l/I_u unlabeled observations.

IV. Unknown Association of the Known Densities with their Labels

Let now two densities $g_1(\cdot)$ and $g_2(\cdot)$ be given. Assume that from the form of the densities one cannot infer whether $g_1(\cdot) = f_1(\cdot)$ and $g_2(\cdot) = f_2(\cdot)$ or if $g_1(\cdot) = f_2(\cdot)$ and $g_2(\cdot) = f_1(\cdot)$. Let the probability η_0 of observing a sample of Class 1 be unknown and call ζ_0 the mixing parameter associated with $g_1(\cdot)$. Define a random variable Z by

$$\begin{aligned} Z &= 1 \quad \text{if} \quad g_1(\cdot) = f_1(\cdot),\ g_2(\cdot) = f_2(\cdot) \ \text{ and } \ \zeta_0 = \eta_0 \\ Z &= 2 \quad \text{if} \quad g_1(\cdot) = f_2(\cdot),\ g_2(\cdot) = f_1(\cdot) \ \text{ and } \ \overline{\zeta}_0 = \eta_0 \end{aligned}$$

and let $\Pr\{Z = 1\} = \Pr\{Z = 2\} = 1/2$ to indicate that the forms of $g_1(\cdot)$ and $g_2(\cdot)$ do not help in deciding whether $g_1(\cdot) = f_1(\cdot)$ and $g_2(\cdot) = f_2(\cdot)$ or if the opposite holds ($g_1(\cdot) = f_2(\cdot)$ and $g_2(\cdot) = f_1(\cdot)$). It is worth noting that the conditional distribution of the unlabeled samples given ζ does not depend on Z. We observe a training set Q composed of l labeled samples and u unlabeled samples, independently distributed as described in the previous section. We are interested in the behavior of $R(l,u)$, the risk of a Bayes classifier with respect to a smooth prior $h(\zeta)$. Note that here we put a prior on ζ, while in Section III the prior was on η. The notation ζ indicates the random variable (or the dummy variable in integration and differentiation) and ζ_0 denotes the actual value of the mixing parameter associated with $g_1(\cdot)$.

One could be tempted to extend the results of Theorem 1 to the current framework and write $\Delta R(l,u) \sim c[uI_u + lI_l]^{-1}$. This conclusion holds true when $u = O(l)$, but care must be taken when the ratio l/u converges to zero as the sample size grows to infinity. In particular, one cannot construct a useful Bayesian classifier relying only on unlabeled samples, i.e., $R(0,u) = 1/2 \ \forall u$.

We have addressed a particular case of this problem in a previous paper [2], where we have proved that, when the training set contains an infinite number of unlabeled samples and l labeled samples, the probability of error $R(l,\infty)$ converges to the Bayes risk R^* exponentially fast in the number of labeled observations, with exponent

$$D = -\log \left\{ 2\sqrt{\eta\overline{\eta}} \int \sqrt{f_1(x) f_2(x)} \, dx \right\}.$$

Here we analyze the dependence of the additional risk $\Delta R(l,u)$ on the number of labeled and unlabeled observations for finite sample size. We consider the large sample behavior of $\Delta R(l,u)$ and we assume that there exists a small constant $\epsilon > 0$ such that $l^{3+\epsilon}u^{-1} \to 0$ as $l + u \to \infty$, the extension of the result to the case $l^{3+\epsilon}u^{-1} \to \infty$, $l = o(u)$ being straightforward. The assumption that $l^{3+\epsilon}u^{-1} \to 0$ is consistent with the basic motivation of the analysis: we analyze cases where the unlabeled samples are cheap and easily available, while the labeled samples are expensive or hard to obtain.

We first write test (6) in a form that simplifies the subsequent analysis. The conditional probability of the event $\{\theta_0 = k\}$ given the training set Q and the new sample X_0 can be rewritten using Bayes theorem as

$$\Pr\{\theta_0 = k | X_0, Q\} = f(X_0 | \theta_0 = k, Q) \frac{\Pr\{\theta_0 = k | Q\}}{f(X_0 | Q)}, \quad k = 1,2. \tag{11}$$

Then the ratio in test (6) can be expressed as

$$\frac{\Pr\{\theta_0 = 1 | X_0, Q\}}{\Pr\{\theta_0 = 2 | X_0, Q\}} = \frac{f(X_0 | \theta_0 = 1, Q) \Pr\{\theta_0 = 1 | Q\}}{f(X_0 | \theta_0 = 2, Q) \Pr\{\theta_0 = 2 | Q\}}. \tag{12}$$

Theorem 2 below is stated in terms of the underlying densities $f_1(\cdot)$ and $f_2(\cdot)$ and addresses the functional dependence of $R(l,u)$ on the number of labeled samples l and of unlabeled samples u for the case $l^{3+\epsilon}u^{-1} \to 0$. In the derivation, let s_η denote the surface of equation $\eta f_1(x) - \overline{\eta} f_2(x) = 0$, let

∇ denote the gradient, let $\|\cdot\|$ be the Euclidean norm, and assume that the integral

$$\int_{s_\eta} \frac{(f_1(x)+f_2(x))^2}{\|\nabla[\eta f_1(x) - \overline{\eta} f_2(x)]\|}\, ds_\eta(x)$$

exists finite and is twice differentiable with respect to η for almost all η, with the possible exception of a finite number of values. Call the Ξ the exceptional subset of $[0,1]$.

Theorem 2: Let the densities $f_1(\cdot)$ and $f_2(\cdot)$ be three times continuously differentiable, with identical support sets. Let the prior $h(\cdot)$ be four times continuously differentiable. Assume that the actual value of the mixing parameter η_0 does not belong to the above defined exceptional set Ξ, and let $0 < \eta_0 < 1$. If $l^{3+\epsilon}u^{-1} \to 0$, the probability of error of the Bayesian test (6) satisfies

$$R(l,u) - R^* = O\left(\frac{1}{u}\right) + \exp\{-Dl + o(l)\} \qquad (13)$$

where the exponent is given by

$$D = -\log\left\{2\sqrt{\eta\overline{\eta}}\int\sqrt{f_1(x)f_2(x)}\,dx\right\}$$

and

$$-\log\int\sqrt{f_1(x)f_2(x)}\,dx$$

is the Bhattacharyya distance between the densities $f_1(\cdot)$ and $f_2(\cdot)$ [29].

We can give here a simple interpretation of the theorem. If the number of labeled samples is small compared to the number of unlabeled samples, $\Delta R(l,u)$, the difference between the risk $R(l,u)$ of the test and the Bayes risk R^* is the sum of two terms.

The first term (which, as seen in the proof, is related to the uncertainty in identifying the correct decision regions or, equivalently, the mixing parameter η_0) depends, to the first order only, on the number u of unlabeled samples in the training set.

The second term reflects the uncertainty in labeling the densities $g_1(\cdot)$ and $g_2(\cdot)$. Asymptotically, the probability of labeling the densities incorrectly depends on the labeled samples alone and converges exponentially fast to zero in the number l of labeled observations in the training set.

We prove the theorem as follows. We first expand the ratio (12) of the optimal Bayesian solution in terms of the conditional posterior expectation of the mixing parameter ζ given $Z = 1$ and $Z = 2$ and of the conditional probabilities of the events $\{Z = 1\}$ and $\{Z = 2\}$ given the training set Q. Then we construct a second test that uses the maximum-likelihood estimator of ζ instead of the posterior expectation, and claim that the first-order terms of $\Delta R(l,u)$ for the Bayesian solution and for the new test are equal, delaying the proof of the statement to Lemma 4 at the end of the section. We show that the test based on the maximum-likelihood estimator is equal to a two-stage procedure. In the first stage, the observation space is partitioned in two decision regions using the maximum-likelihood estimator of the mixing parameter ζ based on the unlabeled samples alone. In the second stage, the decision regions are labeled by deciding that $Z = 1$ if $\Pr\{Z = 1|Q\} > \Pr\{Z = 2|Q\}$ and by deciding that $Z = 2$ otherwise, and the new sample X_0 is classified according to the decision region in which it falls.

In Lemma 2 we show that the probability that the decision regions are labeled incorrectly depends, to the first order only, on the number of labeled samples in the training set, and converges exponentially fast to zero.

In Lemma 3 we show that the test that uses the maximum-likelihood estimator of η to recover the decision regions has probability of error equal to the Bayes risk R^* plus a term of order $O(u^{-1})$.

Lemma 4 finally addresses the claim that the additional probabilities of error of the described procedure and of the Bayesian solution are asymptotically equivalent, in the sense that their ratio converges to one.

Proof: The numerator and the denominator of the ratio (12) can be rewritten by conditioning on Z and taking expectations

$$\begin{aligned} &f(X_0|\theta_0 = k, Q)\Pr\{\theta_0 = k|Q\} \\ &\qquad = E_Z[f(X_0|\theta_0 = k, Q, Z)\Pr\{\theta_0 = k|Q, Z\} \end{aligned}$$

for $k = 1, 2$. When $Z = 1$ and $\theta_0 = 1$, the density of X_0 is $g_1(\cdot)$, when $\theta_0 = 1$ and $Z = 2$, X_0 is distributed according to $g_2(\cdot)$, and similarly we can write the conditional densities for the other combinations of θ_0 and Z. Thus the conditional distribution of X_0 given $Z = k$ and $\theta_0 = m$ is independent of the training set Q. The numerator of (12) becomes

$$\begin{aligned} f(X_0|\theta_0 = 1, Q)&\Pr\{\theta_0 = 1|Q\} \\ &= E_Z[f(X_0|\theta_0 = 1, Q, Z)\Pr\{\theta_0 = 1|Q, Z\}] \\ &= g_1(X_0)\Pr\{\theta_0 = 1|Q, Z = 1\}\Pr\{Z = 1|Q\} \\ &\quad + g_2(X_0)\Pr\{\theta_0 = 1|Q, Z = 2\}\Pr\{Z = 2|Q\} \\ &= g_1(X_0)E_h[\zeta|Q, Z = 1]\Pr\{Z = 1|Q\} \\ &\quad + g_2(X_0)E_h[\overline{\zeta}|Q, Z = 2]\Pr\{Z = 2|Q\} \end{aligned}$$

and, similarly, the denominator becomes

$$\begin{aligned} f(X_0|\theta_0 = 2, Q)&\Pr\{\theta_0 = 2|Q\} \\ &= g_1(X_0)E_h[\zeta|Q, Z = 2]\Pr\{Z = 2|Q\} \\ &\quad + g_2(X_0)E_h[\overline{\zeta}|Q, Z = 1]\Pr\{Z = 1|Q\}. \end{aligned}$$

To simplify the subsequent expressions, define $a \triangleq \Pr\{Z = 1|Q\}$, let b and c denote the conditional expectations of ζ given Q and $Z = 1$, and given Q and $Z = 2$, respectively, i.e.,

$$b \triangleq E_h[\zeta|Q, Z = 1]$$
$$c \triangleq E_h[\zeta|Q, Z = 2]$$

and let $\overline{a} \triangleq 1 - a$, $\overline{b} \triangleq 1 - b$, and $\overline{c} \triangleq 1 - c$. The ratio (12) then becomes

$$\frac{\Pr\{\theta_0 = 1|X_0, Q\}}{\Pr\{\theta_0 = 2|X_0, Q\}} = \frac{abg_1(X_0) + \overline{a}\,\overline{c}g_2(X_0)}{\overline{a}cg_1(X_0) + a\overline{b}g_2(X_0)}. \qquad (14)$$

Let now $\hat{\zeta}^{(u)}$ denote the maximum-likelihood estimator of ζ based on the unlabeled samples. We show in Appendix A2b) that the smoothness of $h(\cdot)$ and the assumption that

$l^{3+\epsilon}u^{-1} \to 0$ imply that $u^{2/3}(b-c) \to 0, u^{2/3}(b-\hat{\zeta}^{(u)}) \to 0$, and $u^{2/3}(c-\hat{\zeta}^{(u)}) \to 0$. It follows that all Bayes tests with respect to a three times continuously differentiable prior are asymptotically equivalent to the test

$$\begin{aligned} &\text{decide } \hat{\theta}_0 = 1 \quad \text{if } \frac{a\hat{\zeta}^{(u)}g_1(X_0) + \overline{a}\overline{\hat{\zeta}}^{(u)}g_2(X_0)}{\overline{a}\hat{\zeta}^{(u)}g_1(X_0) + a\overline{\hat{\zeta}}^{(u)}g_2(X_0)} > 1 \\ &\text{decide } \hat{\theta}_0 = 2 \quad \text{if } \frac{a\hat{\zeta}^{(u)}g_1(X_0) + \overline{a}\overline{\hat{\zeta}}^{(u)}g_2(X_0)}{\overline{a}\hat{\zeta}^{(u)}g_1(X_0) + a\overline{\hat{\zeta}}^{(u)}g_2(X_0)} < 1 \end{aligned} \tag{15}$$

in the sense that the ratio of the difference $R(l,u) - R^*$ for test (15) and of the corresponding difference for the Bayesian solution based on the ratio (14) converges to one as the training set size goes to infinity. A formal proof of this statement is given in Lemma 4, at the end of this section. With some algebra, test (15) can be rewritten as

$$\begin{aligned} &\text{decide } \hat{\theta}_0 = 1 \quad \text{if } \hat{\zeta}^{(u)}g_1(X_0)[a-\overline{a}] > \overline{\hat{\zeta}}^{(u)}g_2(X_0)[a-\overline{a}] \\ &\text{decide } \hat{\theta}_0 = 2 \quad \text{if } \hat{\zeta}^{(u)}g_1(X_0)[a-\overline{a}] < \overline{\hat{\zeta}}^{(u)}g_2(X_0)[a-\overline{a}]. \end{aligned} \tag{16}$$

We can interpret test (16) as a two stage procedure.

Stage 1: Partition the sample space in the regions $\mathcal{X}_1$ and $\mathcal{X}_2$ defined as

$$\begin{aligned} \mathcal{X}_1 &\triangleq \{\hat{\zeta}^{(u)}g_1(x) > \overline{\hat{\zeta}}^{(u)}g_2(x)\} \\ \mathcal{X}_2 &\triangleq \{\hat{\zeta}^{(u)}g_1(x) < \overline{\hat{\zeta}}^{(u)}g_2(x)\} \end{aligned}$$

using the maximum-likelihood estimator of the mixing parameter ζ based on the unlabeled samples alone.

Stage 2: Label the regions $\mathcal{X}_1$ and $\mathcal{X}_2$ by deciding

$$\begin{aligned} &Z = 1 \quad \text{if } \Pr\{Z = 1|Q\} > \Pr\{Z = 2|Q\} \\ &Z = 2 \quad \text{if } \Pr\{Z = 1|Q\} < \Pr\{Z = 2|Q\}. \end{aligned}$$

Let the new sample X_0 be classified with the label of the region in which it falls.

It is easy to see that the interpretation holds. In fact, recall that $a \triangleq \Pr\{Z = 1|Q\}$ and note that the form of test (16) depends on the sign of $a - \overline{a}$: if $a - \overline{a} > 0$, the test is

$$\begin{aligned} &\text{decide } \hat{\theta}_0 = 1 \quad \text{if } \hat{\zeta}^{(u)}g_1(X_0) > \overline{\hat{\zeta}}^{(u)}g_2(X_0) \\ &\text{decide } \hat{\theta}_0 = 2 \quad \text{if } \hat{\zeta}^{(u)}g_1(X_0) < \overline{\hat{\zeta}}^{(u)}g_2(X_0) \end{aligned} \tag{17}$$

and if $a - \overline{a} < 0$, the inequalities are reversed; namely, the test is

$$\begin{aligned} &\text{decide } \hat{\theta}_0 = 1 \quad \text{if } \hat{\zeta}^{(u)}g_1(X_0) < \overline{\hat{\zeta}}^{(u)}g_2(X_0) \\ &\text{decide } \hat{\theta}_0 = 2 \quad \text{if } \hat{\zeta}^{(u)}g_1(X_0) > \overline{\hat{\zeta}}^{(u)}g_2(X_0) \end{aligned} \tag{18}$$

and the interpretation follows immediately.

To evaluate the risk of test (16) we analyze it in terms of the underlying densities $f_1(\cdot)$ and $f_2(\cdot)$. If $Z = 1$, test (17) is actually

$$\begin{aligned} &\text{decide } \hat{\theta}_0 = 1 \quad \text{if } \hat{\eta}^{(u)}f_1(X_0) > \overline{\hat{\eta}}^{(u)}f_2(X_0) \\ &\text{decide } \hat{\theta}_0 = 2 \quad \text{if } \hat{\eta}^{(u)}f_1(X_0) < \overline{\hat{\eta}}^{(u)}f_2(X_0) \end{aligned} \tag{19}$$

and test (18) is actually

$$\begin{aligned} &\text{decide } \hat{\theta}_0 = 1 \quad \text{if } \hat{\eta}^{(u)}f_1(X_0) < \overline{\hat{\eta}}^{(u)}f_2(X_0) \\ &\text{decide } \hat{\theta}_0 = 2 \quad \text{if } \hat{\eta}^{(u)}f_1(X_0) > \overline{\hat{\eta}}^{(u)}f_2(X_0) \end{aligned} \tag{20}$$

while if $Z = 2$ the opposite holds. Note that, as $\hat{\zeta}^{(u)}$ converges to ζ_0, the risk of test (19) converges to the Bayes risk R^*, while the risk of test (20) converges to $1 - R^*$. Thus test (16) is conditionally identical to test (19) given that the decision regions are labeled correctly, and to test (20) otherwise. Let $A = A(l,u)$ be the event corresponding to the incorrect labeling of the decision regions $\mathcal{X}_1$ and $\mathcal{X}_2$. We can then evaluate the risk of test (16) by means of the total probability theorem as

$$\begin{aligned} R(l,u) &= \Pr\{\text{error}|A\}\Pr\{A\} + \Pr\{\text{error}|\overline{A}\}\Pr\{\overline{A}\} \\ &= \Pr\{\text{error}|\overline{A}\} + \Pr\{A\}(\Pr\{\text{error}|A\} \\ &\quad - \Pr\{\text{error}|\overline{A}\}). \end{aligned} \tag{21}$$

The quantity $(\Pr\{\text{error}|A\}–\Pr\{\text{error}|\overline{A}\})$ converges to $1–2R^*$. To calculate $R(lu)$ we must then evaluate $\Pr\{\text{error}|\overline{A}\}$ and $\Pr\{A\}$.

Lemma 2: Under the assumptions of Theorem 2, the probability $\Pr\{A\}$ of labeling incorrectly the decision regions $\mathcal{X}_1$ and $\mathcal{X}_2$ converges exponentially fast to zero in the number of the labeled samples in the training set, the exponent being equal to $-Dl + o(l)$, where

$$D = -\log\left\{2\sqrt{\eta\overline{\eta}}\int\sqrt{f_1(x)f_2(x)}\,dx\right\}.$$

The proof is in Appendix-A3. Thus the first-order term of the probability of using the right or wrong test is a function of the number of labeled samples alone. The following lemma is devoted to the evaluation of $\Pr\{\text{error}|\overline{A}\}$, the conditional probability of error given the correct labeling of the recovered decision regions. Note that, when the decision regions are labeled correctly, test (16) is equal to test (19). From Lemma 2 it follows that $\Pr\{A\}$ is, to the first order, a function of the number l of labeled samples. Test (19) is based on the maximum-likelihood estimator of ζ based on the unlabeled samples only. Thus $\Pr\{\text{error}|\overline{A}\}$ is equal to the unconditional probability of error of test (19).

Lemma 3: Under the assumptions of Theorem 2, the probability of error of test (19) equals R^* plus a term of order $O(1/u)$.

Proof: By inspection one sees that test (19) is formally equal to test (9), where $\hat{\eta}^{(u)}$ has been substituted for $E_h[\eta|Q]$. Then the lemma follows immediately from the cited result by O'Neill.

We can now prove the claim that the additional probabilities of error of the optimal Bayesian solution to the classification problem and of test (15) are asymptotically equivalent. From (21) and Lemma 2 it follows that to prove the claim we only need to address the relation between the conditional probabilities of error of test (19) and of the optimal Bayesian solution given that the decision regions are labeled correctly.

Lemma 4: Under the assumptions of Theorem 2, the difference of the conditional probabilities of error of test (19) and of the Bayesian solution based on ratio (14) given $Z = 1$ and $\{\Pr\{Z = 1|Q\} \geq \Pr\{Z = 2|Q\}\}$ is of order $o(u^{-1})$. Similarly, the difference of the conditional probabilities of error of test (20) and of the Bayesian solution based on ratio (14) given $Z = 2$ and $\{\Pr\{Z = 2|Q\} \geq \Pr\{Z = 1|Q\}\}$ is of order $o(u^{-1})$.

The proof is in Appendix A4.

Upon substituting the results of Lemma 2 and Lemma 3 in (21), the proof of Theorem 2 is completed. □

A. The Example Continued

We now continue the analysis of the example. Recall that $f_1(x) = 2x, f_2(x) = 2(1-x)$ for $x \in [0,1], f_1(x) = f_2(x) = 0$ otherwise, and that the actual value of the mixing parameter is $\eta_0 = 1/2$. The Chernoff exponent for the classification problem is then equal to

$$\begin{aligned} D &= -\log\left\{2\sqrt{\eta_0\bar{\eta}_0}\int\sqrt{f_1(x)f_2(x)}\,dx\right\} \\ &= -\log\left\{2\sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)}\int_0^1\sqrt{2x\cdot 2(1-x)}\,dx\right\} \\ &= -\log\left\{2\int_0^1\sqrt{x-x^2}\,dx\right\} = \log\frac{4}{\pi}. \end{aligned}$$

Thus by combining the results of Theorem 1 and Theorem 2 we obtain

$$\begin{aligned} R(l,u) - R^* &\sim \frac{c}{lI_l + uI_u} + \exp\{-lD + o(l)\} \\ &\sim \frac{1}{4l + \frac{4}{3}u} + \left(\frac{\pi}{4}\right)^{l+o(l)}. \end{aligned} \tag{22}$$

In Section III-A we have seen that, when $l = O(u)$ and $u = O(l)$

$$R(l,u) - R^* \sim \frac{3/4}{3l+u}.$$

From (22) it follows that, if $l = o(u)$ and $\exp\{lD\}u^{-1} \to \infty$, then

$$R(l,u) - R^* \sim \frac{3/4}{3l+u} = \frac{3/4}{u+o(u)} \sim \frac{3}{4u}$$

and, therefore, the asymptotic behavior of the probability of error depends, to first order only, on the number of unlabeled samples. Thus when $l = o(u)$ and $\exp\{lD\}u^{-1} \to \infty$, labeled samples are three times more valuable than unlabeled samples in reducing the probability of error, but their overall contribution is negligible. If $\exp\{lD\}u^{-1} \to 0$, then from (22) it follows that

$$R(l,u) - R^* \sim \exp\{-lD + o(l)\} = \left(\frac{\pi}{4}\right)^{l+o(l)}$$

and the asymptotic behavior of the excess in the probability of error depends, to the first order, on the number of labeled samples alone. Note that labeled samples are very valuable: adding one labeled sample to the training set reduces the excess probability of error by a factor $4/\pi = 1.27324$.

V. Discussion and Conclusions

We have examined two apparently similar situations. In the first case, where the distributions of the two classes of samples are known, the additional probability of error $\Delta R(l,u)$ of any Bayesian solution to a classification problem with respect to a smooth prior is due to errors in estimating the true value of the mixing parameter η_0, which result in errors in evaluating the boundaries of the decision regions. We have shown that

$$\Delta R(l,u) = R(l,u) - R^* \sim \frac{c}{lI_l + uI_u}$$

where

$$I_l = \frac{1}{\eta_0\bar{\eta}_0} \qquad I_u = \int\frac{[f_1(x) - f_2(x)]^2}{\eta_0 f_1(x) + \bar{\eta}_0 f_2(x)}\,dx$$

and c is a constant that depends only on the densities of the two classes of samples and on the mixing parameter but not on the choice of the prior $h(\cdot)$.

Labeled and unlabeled samples play a very similar role in the construction of a classifier. The labeled samples are I_l/I_u times more valuable than the unlabeled samples in reducing the additional risk $\Delta R(l,u)$. However, a decision rule can be constructed using unlabeled samples alone.

The main result of this work is the analysis of the classification problem under the assumption that two densities $g_1(\cdot)$ and $g_2(\cdot)$ are given but it is not known which distribution is $f_1(\cdot)$, the density of the samples of Class 1, and which distribution is $f_2(\cdot)$. We have defined a random variable Z, where $Z = 1$ if $g_1(\cdot) = f_1(\cdot)$ and $g_2(\cdot) = f_2(\cdot)$, and $Z = 2$ otherwise, and we have assumed that $\Pr\{Z = 1\} = \Pr\{Z = 2\} = 1/2$. Within this framework, not only the boundaries but also the labels of the decision regions must be inferred from the training set. We need labeled samples to construct a classification rule, since unlabeled observations are independent of the random variable Z. We have shown that, if $l^{3+\epsilon}u^{-1} \to 0$ as $(l+u) \to \infty$, the probability of error of any Bayesian classifier with respect to a smooth prior satisfies

$$R(l,u) - R^* = O\left(\frac{1}{u}\right) + \exp\{-lD + o(l)\} \tag{23}$$

where

$$D = -\log\left\{2\sqrt{\eta\bar{\eta}}\int\sqrt{f_1(x)f_2(x)}\,dx\right\}$$

and

$$-\log\left\{\int\sqrt{f_1(x)f_2(x)}\,dx\right\}$$

is the Bhattacharyya distance between the densities $f_1(\cdot)$ and $f_2(\cdot)$. The first term on the right-hand side of (23) is due to the approximation in partitioning the sample space and depends to the first order only on the number of unlabeled samples. The second term is due to the probability of error in labeling the decision regions, or, equivalently, in deciding whether $Z = 1$ or $Z = 2$. The probability of labeling incorrectly the estimated decision regions converges exponentially to zero in the number of labeled samples.

The behavior of the risk can be interpreted as follows: if the number of unlabeled samples u grows faster than $\exp\{Dl\}$, the additional probability of error is asymptotically equivalent to the probability of labeling the decision regions incorrectly, namely $\Delta R(l,u) \sim \exp\{-Dl\}$. On the other hand, if $u \exp\{-Dl\} \to 0$ and $l = o(u)$, $\Delta R(l,u)$ is determined by the error in estimating η_0 from the data, and is asymptotically equivalent to $c(I_u(\eta_0)u)^{-1}$. We conclude that in the second case, labeled samples are necessary and exponentially more valuable than unlabeled samples in constructing a classification rule.

Appendix

The appendix is devoted to the details of the proofs of the supporting lemmas.

A1. Proof of Lemma 1

Lemma 1 states that the ratio used in the Bayesian solution to the classification problem can be expressed in terms of the underlying densities $f_1(\cdot)$ and $f_2(\cdot)$ and of the posterior expectation of the mixing parameter η as

$$\frac{\Pr\{\theta_0 = 1|X_0, Q\}}{\Pr\{\theta_0 = 2|X_0, Q\}} = \frac{f_1(X_0)E_h[\eta|Q]}{f_2(X_0)E_h[\overline{\eta}|Q]}$$

thus allowing us to reduce the Bayesian classification problem to the framework of O'Neill's theorem [21, Theorem 1]. To prove the lemma, expand the conditional probability that X_0 has classification 1 given X_0 and the training set Q as

$$\begin{aligned}\Pr\{\theta_0 = 1|X_0, Q\} &= E_h[\Pr\{\theta_0 = 1|X_0, Q, \eta\}]\\ &= \int_0^1 \Pr\{\theta_0 = 1|X_0, Q, \eta\} f(\eta|X_0, Q)\,d\eta.\end{aligned}$$

The conditional probability that $\theta_0 = 1$ given η and X_0 is independent of Q; consequently

$$\Pr\{\theta_0 = 1|X_0, Q, \eta\} = \Pr\{\theta_0 = 1|X_0, \eta\}.$$

This last conditional probability is easily evaluated using Bayes theorem,

$$\Pr\{\theta_0 = 1|X_0, \eta\} = \frac{\eta f_1(X_0)}{\eta f_1(X_0) + \overline{\eta} f_2(X_0)}.$$

The conditional density of η given X_0 and Q can be written by invoking once more Bayes theorem to obtain

$$f(\eta|X_0, Q) = f(X_0, Q|\eta)h(\eta)/f(X_0, Q)$$

where $f(X_0, Q)$ is independent of η, being equal to the expectation of $f(X_0, Q|\eta)$. Note now that the samples in the augmented training set $\{X_0, Q\}$ are independent given the mixing parameter η, thus

$$\begin{aligned}f(X_0, Q|\eta) &= f(X_0|\eta)f(Q|\eta)\\ &= [\eta f_1(X_0) + \overline{\eta} f_2(X_0)]f(Q|\eta).\end{aligned}$$

We can now rewrite the conditional probability that $\theta_0 = 1$ given X_0 and Q as

$$\begin{aligned}\Pr\{\theta_0 = 1|X_0, Q\} &= \int_0^1 \frac{\eta f_1(X_0)}{\eta f_1(X_0) + \overline{\eta} f_2(X_0)}\\ &\quad \cdot \frac{[\eta f_1(X_0) + \overline{\eta} f_2(X_0)]f(Q|\eta)}{f(X_0, Q)} h(\eta)\,d\eta\\ &= \frac{f_1(X_0)}{f(X_0, Q)} \int_0^1 \eta f(Q|\eta)h(\eta)\,d\eta.\end{aligned}$$

Multiplying and dividing the integrand by $f(Q)$, and applying Bayes theorem yields

$$\begin{aligned}\Pr\{\theta_0 = 1|X_0, Q\} &= \frac{f_1(X_0)}{f(X_0, Q)} \int_0^1 \eta f(Q|\eta)h(\eta)\frac{f(Q)}{f(Q)}\,d\eta\\ &= f_1(X_0)\frac{f(Q)}{f(X_0, Q)} \int_0^1 \eta \frac{f(Q|\eta)h(\eta)}{f(Q)}\,d\eta.\\ &= f_1(X_0)\frac{f(Q)}{f(X_0, Q)} \int_0^1 \eta f(\eta|Q)\,d\eta\\ &= f_1(X_0)\frac{f(Q)}{f(X_0, Q)} E_h[\eta|Q] \qquad (24)\end{aligned}$$

$$\Pr\{\theta_0 = 2|X_0, Q\} = f_2(X_0)\frac{f(Q)}{f(X_0, Q)} E_h[\overline{\eta}|Q]. \qquad (25)$$

The term $f(Q)/f(X_0, Q)$ appears in both (24) and (25); thus it cancels in the ratio of the conditional probabilities. We conclude that

$$\frac{\Pr\{\theta_0 = 1|X_0, Q\}}{\Pr\{\theta_0 = 2|X_0, Q\}} = \frac{f_1(X_0)E_h[\eta|Q]}{f_2(X_0)E_h[\overline{\eta}|Q]}$$

thus completing the proof. □

A2. Bayes Estimators and Maximum-Likelihood Estimators of the Mixing Parameter

The main step in the proof of Theorem 2 is the approximation of Bayesian solution based on the ratio (14) by means of test (15). By inspection one sees that the structure of the two tests are analogous; the Bayesian solution is based on b and c, the conditional posterior expectations of ζ given $Z = 1$ and $Z = 2$, respectively; test (15) instead is based on $\hat{\zeta}^{(u)}$, the value of ζ maximizing the likelihood of the unlabeled samples.

A.2a) Posterior Mean and Maximum Likelihood Estimator: We address the problem of the relation between the maximum-likelihood estimator of ζ_0, denoted by $\hat{\zeta}$, and Bayes estimators with respect to the squared error loss function, with smooth priors $h(\cdot)$. We require that $h(\cdot)$ be four times continuously differentiable. Bayes estimators with respect to the squared-error loss function are the conditional expectation of the parameter given the data, and can be written invoking Bayes theorem as

$$\begin{aligned}E_h[\zeta|Q] &= \int_0^1 \zeta f(\zeta|Q)\,d\zeta = \int_0^1 \zeta f(Q|\zeta)\frac{h(\zeta)}{f(Q)}\,d\zeta\\ &= \frac{\int_0^1 \zeta f(Q|\zeta)h(\zeta)\,d\zeta}{\int_0^1 f(Q|\zeta)h(\zeta)\,d\zeta}. \qquad (26)\end{aligned}$$

If the likelihood of the training set is unimodal, as it is the case when we calculate the quantities $b \triangleq E[\zeta|Q, Z = 1]$ and $c \triangleq E[\zeta|Q, Z = 2]$, (26) can be evaluated using Laplace's method of integration. Following, for instance, the approach of Kass, Tierney, and Kadane (see [13, Theorem 4]), we conclude that

$$E[\zeta|Q, Z = k] = \hat{\zeta} + O\left(\frac{1}{n}\right) \tag{27}$$

where $n = l + u$ is the training set size and $\hat{\zeta}$ is the value maximizing the likelihood $f(\zeta|Q)$.

A2b) The Maximum-Likelihood Estimators $\hat{\zeta}$ and $\hat{\zeta}^{(u)}$: The purpose of this section is to prove the claim that $u^{2/3}(b - \hat{\zeta}^{(u)}) \to 0$ and $u^{2/3}(c - \hat{\zeta}^{(u)}) \to 0$, with the aid of (27), where $\hat{\zeta}^{(u)}$ is the maximum-likelihood estimator of the mixing parameter based on the unlabeled samples alone. We analyze in detail the behavior of b, the other case being analogous. From (27) it follows that

$$u^{2/3}(b - \hat{\zeta}^{(u)}) = u^{2/3}(\hat{\zeta} - \hat{\zeta}^{(u)}) + O\left(u^{-1/3}\right).$$

Thus to prove the claim we only need to bound from above (in probability) the term $|\hat{\zeta} - \hat{\zeta}^{(u)}|$. The average log-likelihood of the training set can be written as

$$u^{-1}L(Q;\zeta) = [u^{-1}L(Q_u;\zeta) + lu^{-1}(l^{-1}L(Q_l;\zeta))].$$

We want to compare the solutions of $(d/d\zeta)u^{-1}L(Q;\zeta) = 0$ and of $(d/d\zeta)u^{-1}L(Q_u;\zeta) = 0$. Note that $\log\{\zeta g_1(x) + \overline{\zeta} g_2(x)\}$ is a strictly concave function of ζ for fixed value of x. The average log-likelihood of the unlabeled samples is therefore strictly concave, being the average of strictly concave functions, and the maximizing value $\hat{\zeta}^{(u)}$ is therefore unique. The same holds for the conditional log-likelihood of the labeled samples given $Z = 1$, i.e.,

$$l^{-1}L(Q_l;\zeta|Z = 1) = l^{-1}\sum_{i=1}^{l} \log\{\zeta_{\theta_i} g_{\theta_i}(X_i)\}$$

where, as usual, $\zeta_{\theta_i} = \zeta$ if $\theta_i = 1, \zeta_{\theta_i} = \overline{\zeta}$ if $\theta_i = 2$. Thus the conditional log-likelihood of the entire training set is a strictly concave function of ζ and has a unique point of maximum $\hat{\zeta}^{(u)}$. We can therefore reduce our analysis of $|\hat{\zeta} - \hat{\zeta}^{(u)}|$ to the study of the roots of the equations

$$u^{-1}(d/d\zeta)L(Q_u;\zeta) = 0$$

and

$$u^{-1}(d/d\zeta)L(Q;\zeta;Z = 1) = 0.$$

Lemma 5: There exist strictly positive constants M and D_ζ such that

$$-\lim_{u\to\infty}\frac{1}{u}\log\Pr\{|\hat{\zeta} - \hat{\zeta}^{(u)}| > M\frac{l}{u}\} \geq D_\zeta.$$

To find an upper bound in probability to $|\hat{\zeta} - \hat{\zeta}^{(u)}|$, we use a worst case approach: we bound in probability from below the absolute value of the derivative of the log-likelihood of the unlabeled samples and from above the absolute value of the derivative of the log-likelihood of the labeled samples, and we use the bounds to evaluate the point of maximum of the log-likelihood of the entire training set.

In the neighborhood of $\hat{\zeta}^{(u)}$ the log-likelihood of the unlabeled samples can be expanded in Taylor series as

$$\begin{aligned}\frac{1}{u}L(Q_u;\zeta) = {}& \frac{1}{u}L(Q_u;\hat{\zeta}^{(u)}) \\ &+ \frac{1}{u}\frac{d}{d\zeta}L(Q_u;\zeta)\bigg|_{\hat{\zeta}^{(u)}}(\zeta - \hat{\zeta}^{(u)}) \\ &+ \frac{1}{u}\frac{d^2}{d\zeta^2}L(Q_u;\zeta)\bigg|_{\hat{\zeta}^{(u)}}\frac{(\zeta - \hat{\zeta}^{(u)})^2}{2} \\ &+ \frac{1}{u}\frac{d^3}{d\zeta^3}L(Q_u;\zeta)\bigg|_{\tilde{\zeta}}\frac{(\zeta - \hat{\zeta}^{(u)})^3}{6}\end{aligned}$$

for some $\tilde{\zeta}$ of the form $\lambda\hat{\zeta}^{(u)} + \overline{\lambda}\zeta, \lambda \in [0,1]$. The first derivative of $L(Q_u;\zeta)$ evaluated at $\hat{\zeta}^{(u)}$ is equal to zero, since $\hat{\zeta}^{(u)}$ is the point of maximum. Consider then

$$\begin{aligned}\frac{1}{u}\frac{d^2}{d\zeta^2}L(Q_u;\zeta)\bigg|_{\hat{\zeta}^{(u)}} &= \frac{1}{u}\sum_{j=1}^{u}\frac{d^2}{d\zeta^2}\log\{\zeta g_1(X'_j) + \overline{\zeta} g_2(X'_j)]\bigg|_{\hat{\zeta}^{(u)}} \\ &= -\frac{1}{u}\sum_{j=1}^{u}\frac{[g_1(X'_j) - g_2(X'_j)]^2}{[\hat{\zeta}^{(u)}g_1(X'_j) + \overline{\hat{\zeta}}^{(u)}g_2(X'_j)]^2}.\end{aligned} \tag{28}$$

The right-hand side of (28) is the average of i.i.d. random variables, the expectation of which is finite and negative for all $\hat{\zeta}^{(u)}$ close to ζ_0. We now show that there exists an interval centered on ζ_0 where the average (28) is uniformly smaller than $-\epsilon$ with conditional probability converging exponentially fast to 1 in u given the event $\{\zeta_0 - \delta \leq \hat{\zeta}^{(u)} \leq \zeta_0 + \delta\}$, for some appropriate choice of $\delta > 0$. Define the event

$$A_1(\epsilon,\delta) \triangleq \left\{\exists\zeta \in (\zeta_0 - 2\delta, \zeta_0 + 2\delta): -\frac{1}{u}\sum_{j=1}^{u}\frac{[g_1(X'_j) - g_2(X'_j)]^2}{[\zeta g_1(X'_j) + \overline{\zeta}_0 g_2(X'_j)]^2} > -\epsilon\right\}.$$

Lemma 6: There exist strictly positive quantities D_ϵ, and ϵ, δ such that

$$-\lim_{n\to\infty}\frac{1}{u}\log\Pr\{A_1(\epsilon,\delta)|\,|\hat{\zeta}^{(u)} - \zeta_0| < \delta\} = D_\epsilon. \tag{29}$$

Proof: Fix a small quantity $\epsilon > 0$ such that $I_u(\zeta_0) > 2\epsilon$, where $I_u(\zeta_0)$ is the Fisher information of the unlabeled samples. The average second derivatives of the log-likelihood of the unlabeled samples evaluated at ζ_0 converges almost surely to minus the Fisher information. From Cramér's theorem (see, for instance, Dembo and Zeitouni [6]) it follows that there exist constants D_ϵ and u_0 greater than zero such that, for all $u > u_0$

$$\Pr\left\{-\frac{1}{u}\sum_{j=1}^{u}\frac{(g_1(X'_j) - g_2(X'_j))^2}{[\zeta_0 g_1(X'_j) + \overline{\zeta}_0 g_2(X'_j)]^2} > -2\epsilon\right\} < \exp\{-uD_\epsilon\}.$$

Use the intermediate value theorem to rewrite the second derivative of the log-likelihood as

$$\begin{aligned}\frac{1}{u}\frac{d^2}{d\zeta^2}L(Q_u;\zeta)\Big|_{\zeta} \\ = \frac{1}{u}\frac{d^2}{d\zeta^2}L(Q_u;\zeta)\Big|_{\zeta_0} + \frac{1}{u}\frac{d^3}{d\zeta^3}L(Q_u;\zeta)\Big|_{\tilde{\zeta}}(\zeta-\zeta_0)\end{aligned} \tag{30}$$

for $\tilde{\zeta} = \alpha\zeta_0 + \overline{\alpha}\zeta$, some appropriate $\alpha \in [0,1]$. The third derivative of the average log-likelihood with respect to ζ can be bounded as

$$\begin{aligned}\frac{1}{u}\frac{d^3}{d\zeta^3}L(Q_u;\zeta) &= \frac{1}{u}\sum_{j=1}^{u}\frac{[g_1(X'_j)-g_2(X'_j)]^3}{[\zeta g_1(X'_j)+\overline{\zeta}g_2(X'_j)]^3} \\ &\le \frac{1}{\min\{\zeta^3,\overline{\zeta}^3\}}.\end{aligned} \tag{31}$$

Then we can bound from above the second derivative (30) as follows:

$$\begin{aligned}\frac{1}{u}\frac{d^2}{d\zeta^2}L(Q_u;\zeta)\Big|_{\zeta} \\ \le \frac{1}{u}\frac{d^2}{d\zeta^2}L(Q_u;\zeta)\Big|_{\zeta_0} \\ + |\zeta-\zeta_0| \sup_{\substack{\tilde{\zeta}=a\zeta_0+\overline{a}\zeta,\\ a\in[0,1]}} \left|\frac{1}{u}\frac{d^3}{d\zeta^3}L(Q_u;\zeta)|_{\tilde{\zeta}}\right| \\ \le \frac{1}{u}\frac{d^2}{d\zeta^2}L(Q_u;\zeta)\Big|_{\zeta_0} \\ + \frac{|\zeta-\zeta_0|}{\min\{(\zeta_0-|\zeta-\zeta_0|)^3,(1-\zeta_0-|\zeta-\zeta_0|)^3\}}\end{aligned}$$

where the last inequality is a direct consequence of upper bound (31). Conditional on the event

$$\{(1/u)(d^2/d\zeta^2)L(Q_u;\zeta)|_{\zeta_0} \le 2\epsilon\}$$

the second derivative $(1/u)(d^2/d\zeta^2)L(Q_u;\zeta)$ is less than or equal to ϵ for every ζ in the set

$$\mathcal{Z} \triangleq \left\{\frac{|\zeta-\zeta_0|}{\min\{(\zeta_0-|\zeta-\zeta_0|)^3,(1-\zeta_0-|\zeta-\zeta|)^3\}1} \le \epsilon\right\}.$$

Let $2\delta'$ be the supremum of $|\zeta-\zeta_0|$ over the set $\mathcal{Z}$ Recall the assumption that there exists a symmetric open interval centered on η_0 all contained in the parameter space $[0,1]$, let $2\delta''$ be the half width of this symmetric interval and define $\delta = \min\{\delta',\delta''\}$. Then

$$\lim_{u\to\infty}(1/u)\log\Pr\{A_1(\epsilon,\delta)|\,|\hat{\zeta}^{(u)}-\zeta_0|<\delta\} = -D_\epsilon.$$

Note that a suitable choice for δ is $\epsilon\min\{(\zeta_0)^3,(1-\zeta_0)^3\}$.□

Lemma 7: There exists a positive constant D_u such that

$$\lim_{u\to\infty}\frac{1}{u}\log\Pr\{|\hat{\zeta}^{(u)}-\zeta_0|\}>\delta = -D_u,$$

where δ is the quantity defined in the previous lemma.

Proof: The log-likelihood of the unlabeled samples is a strictly concave function of ζ for all values of x. Consequently, the average log-likelihood of the unlabeled samples in the training set, $u^{-1}L(Q_u;\zeta)$, is a strictly concave function, being the average of strictly concave functions, and thus it has a unique maximum. The average log-likelihood converges to its expectation, which has a unique maximum at the true value of the parameter ζ_0. This is easily seen, as

$$\begin{aligned}&\int[\zeta_0 g_1(x)+\overline{\zeta}_0 g_2(x)]\log[\zeta g_1(x)+\overline{\zeta}g_2(x)]\,dx \\ &\quad= \int[\zeta_0 g_1(x)+\overline{\zeta}_0 g_2(x)]\log[\zeta_0 g_1(x)+\overline{\zeta_0}g_2(x)]\,dx \\ &\qquad- \int[\zeta_0 g_1(x)+\overline{\zeta}_0 g_2(x)]\log\frac{\zeta_0 g_1(x)+\overline{\zeta}_0 g_2(x)}{\zeta g_1(x)+\overline{\zeta}g_2(x)}\,dx \\ &\quad= h(\zeta_0 g_1(x)+\overline{\zeta}_0 g_2(x)) - D(\zeta_0 g_1(x)+\overline{\zeta}_0 g_2(x)\| \\ &\qquad\cdot\zeta g_1(x)+\overline{\zeta}g_2(x)) \\ &\quad\le h(\zeta_0 g_1(x)+\overline{\zeta}_0 g_2(x))\end{aligned}$$

where $h(\cdot)$ is the differential entropy and $D(\cdot\|\cdot)$ is the Kullback–Leibler distance, which is greater than zero if the two arguments are different. By inspection one sees that the average log-likelihood of the unlabeled samples is differentiable with respect to ζ. From the differentiability and the concavity of the log-likelihood, it follows that

$$\{\hat{\zeta}^{(u)}<\zeta\} = \left\{\frac{d}{d\zeta'}L(Q_u;\zeta')|_{\zeta}<0\right\}$$

and

$$\{\hat{\zeta}^{(u)}>\zeta\} = \left\{\frac{d}{d\zeta'}L(Q_u;\zeta')|_{\zeta}>0\right\}.$$

It follows that

$$\begin{aligned}\Pr\{|\hat{\zeta}^{(u)}-\zeta_0|>\delta\} &= \Pr\left\{\frac{1}{u}\frac{d}{d\zeta'}L(Q_u;\zeta')\Big|_{\zeta_0-\delta}<0\right\} \\ &\quad+\Pr\left\{\frac{1}{u}\frac{d}{d\zeta'}L(Q_u;\zeta')\Big|_{\zeta_0+\delta}>0\right\}\end{aligned} \tag{32}$$

where the equality holds because the events on the right-hand side are disjoint. We analyze the first probability on the right-hand side of equality (32), the other case being analogous. The derivative of the log-likelihood of the unlabeled samples is given by

$$\frac{1}{u}\frac{d}{d\zeta}L(Q_u;\zeta)\Big|_{\zeta} = \frac{1}{u}\sum_{j=1}^{u}\frac{g_1(X'_j)-g_2(X'_j)}{\zeta g_1(X'_j)+\overline{\zeta}g_2(X'_j)}$$

which is the average of i.i.d. random variables. The expectation of the derivative evaluated at $\zeta' \triangleq \zeta_0-\delta$ is greater than zero: it is easily shown that

$$E\left[\frac{1}{u}\frac{d}{d\zeta}L(Q_u;\zeta)\Big|_{\zeta}\right] = \delta\int\frac{[g_1(x)-g_2(x)]^2}{\zeta' g_1(x)+\overline{\zeta}' g_2(x)}\,dx>0. \tag{33}$$

Then, from Cramér's theorem it follows that

$$\begin{aligned}-\lim_{u\to\infty}\frac{1}{u}\log\Pr\left\{\frac{1}{u}\frac{d}{d\zeta}L(Q_u;\zeta)\Big|_{\zeta_0-\delta}<0\right\} &= \inf_{x\le 0}\Lambda^*(x) \\ &= \Lambda^*(0)\end{aligned}$$

where $\Lambda^*(\cdot)$ is the rate function associated with the large deviations principle for the average of the derivatives of the log-likelihood, and the last equality follows from of Dembo and Zeitouni [6, Lemma 2.2.5]. From the same lemma it also follows that $\Lambda^*(0) = -\inf_{\alpha \leq 0} \Lambda(\alpha)$, where $\Lambda(\cdot)$ is the logarithmic moment generating function of the derivative of the log-likelihood, and is convex. It is easy to see that $\Lambda(\cdot)$ is finite for all negative values of α, being equal to

$$\Lambda(\alpha) = \log \int \exp\left\{\alpha \frac{g_1(x) - g_2(x)}{(\zeta_0 + \delta) g_1(x) + (\overline{\zeta}_0 - \delta) g_2(x)}\right\} \cdot [\zeta_0 g_1(x) + \overline{\zeta}_0 g_2(x)]\, dx.$$

The derivative of $\Lambda(\alpha)$ evaluated at $\alpha = 0$ is strictly positive for $\delta > 0$, and from the convexity of $\Lambda(\cdot)$ and from the equalities $\Lambda(0) = 0, \Lambda(-\infty) = \infty$, we conclude that there exists a negative value of α where the logarithmic moment generating function attains its minimum $-D_u$ and has strictly negative value. □

Define the events $A_2(\epsilon)$ and $A_3(\delta)$ as

$$A_2(\epsilon) \triangleq \left\{\exists \zeta \in (\zeta_0 - 2\delta, \zeta_0 + 2\delta): \left|\frac{1}{u}\frac{d}{d\zeta'} L(Q_u; \zeta')|_\zeta\right| < \epsilon |\zeta - \hat{\zeta}^{(u)}|\right\}$$

$$A_3(\delta) \triangleq \{|\hat{\zeta}^{(u)} - \zeta_0| > \delta\}.$$

The following lemma bounds from above the probability of their union. On the complement of $A_2(\epsilon) \bigcup A_3(\delta)$, the absolute value of the first derivative of the average log-likelihood of the unlabeled samples is bounded from below by $\epsilon|\zeta - \hat{\zeta}^{(u)}|$ for every $\zeta \in (\zeta_0 - 2\delta, \zeta_0 + 2\delta)$.

Lemma 8: There exist positive quantities ϵ, δ and D'_u such that

$$-\lim_{u \to \infty} \frac{1}{u} \log \Pr\{A_1(\epsilon) \bigcup A_2(\delta)\} > D'_u$$

where ϵ is the same quantity defined in Lemma 6, and $\delta = \epsilon \min\{\zeta_0^3, \overline{\zeta}_0^3\}$.

Proof: Lemma 6 states that the probability that the second derivative of the the log-likelihood of the labeled samples evaluated at any point of the interval $(\zeta_0 - 2\delta, \zeta_0 + 2\delta)$ is greater than $-\epsilon$ converges exponentially fast to 0 in the number of unlabeled samples, with exponent D_ζ. The intermediate value theorem applied to the first derivative yields

$$-\lim_{u \to \infty} \log \Pr\left\{\left|\frac{1}{u}\frac{d}{d\zeta'} L(Q_u; \zeta')|_\zeta\right| < \epsilon|\zeta - \hat{\zeta}^{(u)}|\right\} > D'_u. \tag{34}$$

The proof is concluded by applying the union of events bound and Lemma 7. □

If we use the lower bound (34) instead of the actual derivative of the log-likelihood to calculate the value of $\hat{\zeta}$, we overestimate the distance between $\hat{\zeta}^{(u)}$ and $\hat{\zeta}$.

Consider now the behavior of the log-likelihood of the labeled samples. To assess the influence of the unlabeled samples we calculate the maximum of the derivative of the log-likelihood in the set $(\zeta_0 - 2\delta, \zeta_0 + 2\delta)$.

Lemma 9: The first derivative of the log-likelihood of the labeled samples satisfies

$$\sup_{\zeta \in (\zeta_0 - 2\delta, \zeta_0 + 2\delta)} \left|\frac{1}{l}\frac{d}{d\zeta} L(Q_l; \zeta)\right| \leq K \triangleq \frac{1}{[\min(\zeta_0 - 2\delta, 1 - \zeta_0 - 2\delta)]}. \tag{35}$$

Proof: Recall that we are considering the case $Z = 1$, namely, here $g_1(\cdot) = f_1(\cdot), g_2(\cdot) = f_2(\cdot)$ and $\zeta = \eta$. Then

$$\frac{1}{l} L(Q_l; \zeta) = \frac{1}{l} \log\left\{\prod_{i=1}^{l} \zeta_{\theta_i} g_{\theta_i}(X_i)\right\} = \frac{1}{l} \log\left\{\zeta^{l_1} \overline{\zeta}^{l_2} \prod_{i=1}^{l} g_{\theta_i}(X_i)\right\}$$

where l_1 and l_2 are, respectively, the numbers of samples of Class 1 and of samples of Class 2. Taking the absolute value of the derivative with respect to ζ yields

$$\begin{aligned}\left|\frac{1}{l}\frac{d}{d\zeta} L(Q_l; \zeta)\right| &= \left|\frac{1}{l}\frac{d}{d\zeta} \log\left\{\zeta^{l_1} \overline{\zeta}^{l_2} \prod_{i=1}^{l} g_{\theta_i}(X_i)\right\}\right| \\ &= \left|\frac{1}{l}\frac{d}{d\zeta} \log \zeta^{l_1} + \frac{1}{l}\frac{d}{d\zeta} \log|, \overline{\zeta}^{l_2}\right| \\ &= \left|\frac{1}{l}\frac{l_1}{\zeta} - \frac{1}{l}\frac{l_2}{\overline{\zeta}}\right| = \left|\frac{l_1 \overline{\zeta} - l_2 \zeta}{\zeta \overline{\zeta}}\right| \\ &\leq \frac{1}{\min(\zeta, \overline{\zeta})}.\end{aligned} \tag{36}$$

Thus the lemma follows immediately. □

The value of ζ maximizing the likelihood of the training set Q is the solution to the equation

$$\frac{d}{d\zeta}\left(\frac{1}{u} L(Q_u; \zeta) + \frac{l}{u}\frac{1}{l} L(Q_l; \zeta | Z = 1)\right) = 0.$$

If we substitute in this equation the lower bound (34) of the derivative of $u^{-1}L(Q_u; \zeta)$ and the upper bound (35) of the derivative of $l^{-1}L(Q_l; \zeta | Z = 1)$, we overestimate $|\hat{\zeta} - \hat{\zeta}^{(u)}|$. The resulting equation is $(-\epsilon|\zeta - \hat{\zeta}^{(u)}| + K(l/u)) = 0$, the solution of which bounds from above $|\hat{\zeta} - \hat{\zeta}^{(u)}|$. We conclude that there exist constants M and D_ζ such that $|\hat{\zeta} - \hat{\zeta}^{(u)}| < Mlu^{-1}$ with probability greater than $1 - \exp\{-D_\zeta u\}$. We conclude the proof of Lemma 5 by noting that the derivation for the case $Z = 2$ is identical the one just completed. □

A3. Proof of Lemma 2

Lemma 2 states that the probability $\Pr\{A\}$ of labeling incorrectly the recovered decision regions $\mathcal{X}_1$ and $\mathcal{X}_2$ is equal to $\exp\{-Dl + o(l)\}$, where the exponent is

$$D = -\log\left\{2\sqrt{\eta\overline{\eta}} \int \sqrt{f_1(x) f_2(x)}\, dx\right\}.$$

The regions $\mathcal{X}_1$ and $\mathcal{X}_2$ are labeled by deciding that $Z = 1$ if

$$\Pr\{Z = 1 | Q\} > \Pr\{Z = 2 | Q\}$$

and by deciding $Z = 2$ if the inequality is reversed. Thus we now consider the properties of the ratio

$$\frac{\Pr\{Z=1|Q\}}{\Pr\{Z=2|Q\}} = \frac{f(Q|Z=1)\Pr\{Z=1\}/f(Q)}{f(Q|Z=2)\Pr\{Z=2\}/f(Q)} = \frac{f(Q|Z=1)}{f(Q|Z=2)} \tag{37}$$

where the last equality follows from the assumption

$$\Pr\{Z=1\} = \Pr\{Z=2\} = 1/2.$$

Taking conditional expectations yields

$$\frac{f(Q|Z=1)}{f(Q|Z=2)} = \frac{\displaystyle\int_0^1 f(Q_u|\zeta, Z=1) f(Q_l|\zeta, Z=1) h(\zeta)\, d\zeta}{\displaystyle\int_0^1 f(Q_u|\zeta, Z=2) f(Q_l|\zeta, Z=2) h(\zeta)\, d\zeta}.$$

We have previously mentioned that the distribution of the unlabeled samples given ζ is independent of Z. Therefore

$$\frac{f(Q|Z=1)}{f(Q|Z=2)} = \frac{\displaystyle\int_0^1 f(Q_u|\zeta) f(Q_l|\zeta, Z=1) h(\zeta)\, d\zeta}{\displaystyle\int_0^1 f(Q_u|\zeta) f(Q_l|\zeta, Z=2) h(\zeta)\, d\zeta} = \frac{\displaystyle\int_0^1 f(Q_l|\zeta, Z=1) f(\zeta|Q_u)\, d\zeta}{\displaystyle\int_0^1 f(Q_l|\zeta, Z=2) f(\zeta|Q_u)\, d\zeta} \tag{38}$$

where the last step leading to (38) is a simple application of Bayes theorem. The following lemma is needed in the proof.

Lemma 10: If $h(\zeta)$ is three times continuously differentiable on $[0, 1]$, then

$$\int_0^1 f(Q_l|\zeta, Z=1) f(\zeta|Q_u)\, d\zeta = \prod_{i=1}^{l} \hat{\zeta}_{\theta_i}^{(u)} g_{\theta_i}(X_i)\left[1 + O\left(\frac{l}{u}\right)\right] \tag{39}$$

$$\int_0^1 f(Q_l|\zeta, Z=2) f(\zeta|Q_u)\, d\zeta = \prod_{i=1}^{l} \hat{\zeta}_{\overline{\theta}_i}^{(u)} g_{\overline{\theta}_i}(X_i)\left[1 + O\left(\frac{l}{u}\right)\right] \tag{40}$$

where, again, $\hat{\zeta}^{(u)}$ is the maximum-likelihood estimator of ζ based on unlabeled samples alone.

The proof is in Appendix A3a). □

Combining (37), (38), and Lemma 10 with the assumption that $l^{3+\epsilon}u^{-1}$ converges to 0, we conclude that, for l and u large enough

$$\frac{1}{2}\frac{\displaystyle\prod_{i=1}^{l}\hat{\zeta}_{\theta_i}^{(u)} g_{\theta_i}(X_i)}{\displaystyle\prod_{i=1}^{l}\hat{\zeta}_{\overline{\theta}_i}^{(u)} g_{\overline{\theta}_i}(X_i)} < \frac{a}{\overline{a}} < 2\frac{\displaystyle\prod_{i=1}^{l}\hat{\zeta}_{\theta_i}^{(u)} g_{\theta_i}(X_i)}{\displaystyle\prod_{i=1}^{l}\hat{\zeta}_{\overline{\theta}_i}^{(u)} g_{\overline{\theta}_i}(X_i)}.$$

Let now $\Delta\zeta^{(u)} \triangleq \hat{\zeta}^{(u)} - \zeta_0$. It is easy to check that the following inequalities hold:

$$\prod_{i=1}^{l}\hat{\zeta}_{\theta_i}^{(u)} g_{\theta_i}(X_i) \le (1 + c|\Delta\zeta^{(u)}|)^l \prod_{i=1}^{l}\zeta_{\theta_i} g_{\theta_i}(X_i)$$

$$\prod_{i=1}^{l}\hat{\zeta}_{\theta_i}^{(u)} g_{\theta_i}(X_i) \ge (1 - c|\Delta\zeta^{(u)}|)^l \prod_{i=1}^{l}\zeta_{\theta_i} g_{\theta_i}(X_i)$$

for

$$c = \max_{\mathbb{R}}(f_1(x) + f_2(x)) = \max_{\mathbb{R}}(g_1(x) + g_2(x))$$

which is finite by assumption. From the asymptotic normality of $\hat{\eta}^{(u)}$, it follows that $\Delta\zeta^{(u)} = o_p(u^{-1/2+\epsilon})$. From the assumption that $l = O(\sqrt[3+\epsilon]{u})$, it follows that both $(1 - c|\Delta\zeta^{(u)}|)^l$ and $(1 + c|\Delta\zeta^{(u)}|)^l$ converge to one in probability, and that, for u large enough, $(1 + c|\Delta\zeta^{(u)}|)^l > 2$ and $(1 - c|\Delta\zeta^{(u)}|)^l < 1/2$ with negligible probability. Thus we conclude that

$$\Pr\left\{\prod_{i=1}^{l}\frac{\zeta_{\theta_i} g_{\theta_i}(X_i)}{\overline{\zeta}_{\theta_i} g_{\overline{\theta}_i}(X_i)} < \frac{1}{4}\right\} < \Pr\left\{\frac{a}{\overline{a}} < 1\right\} < \Pr\left\{\prod_{i=1}^{l}\frac{\zeta_{\theta_i} g_{\theta_i}(X_i)}{\overline{\zeta}_{\theta_i} g_{\overline{\theta}_i}(X_i)} < 4\right\}$$

$$\Pr\left\{\prod_{i=1}^{l}\frac{\zeta_{\theta_i} g_{\theta_i}(X_i)}{\overline{\zeta}_{\theta_i} g_{\overline{\theta}_i}(X_i)} > 4\right\} < \Pr\left\{\frac{a}{\overline{a}} > 1\right\} < \Pr\left\{\prod_{i=1}^{l}\frac{\zeta_{\theta_i} g_{\theta_i}(X_i)}{\overline{\zeta}_{\theta_i} g_{\overline{\theta}_i}(X_i)} > \frac{1}{4}\right\}. \tag{41}$$

We can now bound the probability of labeling incorrectly the decision regions in terms of the underlying densities, as

$$\Pr\left\{\prod_{i=1}^{l}\frac{\eta_{\theta_i} f_{\theta_i}(X_i)}{\overline{\eta}_{\theta_i} f_{\overline{\theta}_i}(X_i)} < \frac{1}{4}\right\} \le \Pr\{A\} \le \Pr\left\{\prod_{i=1}^{l}\frac{\eta_{\theta_i} f_{\theta_i}(X_i)}{\overline{\eta}_{\theta_i} f_{\overline{\theta}_i}(X_i)} < 4\right\}.$$

But, as we have shown in a previous work ([2, Theorem 2])

$$\lim_{l\to\infty}\frac{1}{l}\log\Pr\left\{\prod_{i=1}^{l}\frac{\eta_{\theta_i} f_{\theta_i}(X_i)}{\overline{\eta}_{\theta_i} f_{\overline{\theta}_i}(X_i)} < \frac{1}{4}\right\} = \lim_{l\to\infty}\frac{1}{l}\log\Pr\left\{\prod_{i=1}^{l}\frac{\eta_{\theta_i} f_{\theta_i}(X_i)}{\overline{\eta}_{\theta_i} f_{\overline{\theta}_i}(X_i)} < 4\right\} = D \tag{42}$$

where

$$D = -\log\left\{2\sqrt{\eta\overline{\eta}}\int_{\mathbb{R}}\sqrt{f_1(x)f_2(x)}\, dx\right\}$$

is the Chernoff exponent for testing $Z = 1$ against $Z = 2$. We conclude the proof of the lemma by combining (41) and (42). □

A3a) Proof of Lemma 10: Write the numerator of (38) as

$$\int f(Q_l|\zeta, Z=1) f(\zeta|Q_u)\, d\zeta = \frac{\displaystyle\int_0^1 f(Q_l|\zeta, Z=1) f(Q_u|\zeta) h(\zeta)\, d\zeta}{f(Q_u)}. \tag{43}$$

Rewrite the numerator of the ratio (43) as

$$\int_0^1 f(Q_l|\zeta, Z=1) f(Q_u|\zeta) h(\zeta)\, d\zeta = \int_0^1 \exp\left\{u\left[\frac{1}{u}L(Q_u;\zeta) + \frac{l}{u}\frac{1}{l}L^{(1)}(Q_l;\zeta)\right]\right\} h(\zeta)\, d\zeta \tag{44}$$

where

$$L^{(1)}(Q_l) = \sum_{i=1}^{l} \log\{\zeta_{\theta_i} g_{\theta_i}(X_i)\}$$

and

$$L^{(2)}(Q_l) = \sum_{i=1}^{l} \log\{\overline{\zeta}_{\theta_i} g_{\overline{\theta}_i}(X_i)\}$$

are, respectively, the conditional log-likelihood of the labeled samples given $Z = 1$ and $Z = 2$. Using Laplace's method for integration (see, for example, [14]), we can write

$$\begin{aligned}\int_0^1 &\exp\left\{u\left[\frac{1}{u}L(Q_u;\zeta) + \frac{l}{u}\frac{1}{l}L^{(1)}(Q_l;\zeta)\right]\right\} h(\zeta)\, d\zeta \\ &= h(\hat{\zeta}) \exp\{L(Q_u;\hat{\zeta}) + L^{(1)}(Q_l;\hat{\zeta})\} \\ &\quad\cdot \left[\frac{2\pi}{u\dfrac{d^2}{d\eta^2}\left.\left[\dfrac{1}{u}L(Q_u;\zeta) + \dfrac{l}{u}\dfrac{1}{l}L^{(1)}(Q_l;\zeta)\right]\right|_{\hat{\zeta}}}\right]^{1/2} \\ &\quad\cdot \left[1 + O\left(\frac{1}{u}\right)\right].\end{aligned} \tag{45}$$

A similar expression holds for

$$f(Q_u) = \int f(Q_u;\zeta) h(\zeta)\, d\zeta$$

$$\begin{aligned}f(Q_u)k &= h(\hat{\zeta}^{(u)}) \exp\{L(Q_u;\hat{\zeta}^{(u)})\} \\ &\quad\cdot \left[\frac{2\pi}{u\dfrac{d^2}{d\zeta^2}\left.L(Q_u;\zeta)\right|_{\hat{\zeta}^{(u)}}}\right]^{1/2} \left[1 + O\left(\frac{1}{u}\right)\right].\end{aligned}$$

Therefore, ratio (43) is equal to

$$\begin{aligned}&\frac{\displaystyle\int f(Q_l|\zeta, Z=1) f(Q_u|\zeta) h(\zeta)\, d\zeta}{f(Q_u)} \\ &= \frac{h(\hat{\zeta})}{h(\hat{\zeta}^{(u)})} \exp\{L(Q_u;\hat{\zeta}) - L(Q_u;\hat{\zeta}^{(u)}) + L(Q_l;\hat{\zeta})\} \\ &\quad\cdot \left[\frac{\dfrac{d^2}{d\zeta^2}\left.\left[\dfrac{1}{u}L(Q_u;\zeta) + \dfrac{l}{u}\dfrac{1}{l}L^{(1)}(Q_l;\zeta)\right]\right|_{\hat{\zeta}}}{\dfrac{d^2}{d\zeta^2}\left.L(Q_u;\zeta)\right|_{\hat{\zeta}^{(u)}}}\right]^{1/2} \\ &\quad\cdot \left[1 + O\left(\frac{1}{u}\right)\right].\end{aligned} \tag{46}$$

Consider the first three factors in (46). From the finiteness of $h(\hat{\zeta}^{(u)})$ and the boundedness of its first derivative, it follows that

$$\frac{h(\hat{\zeta})}{h(\hat{\zeta}^{(u)})} = \frac{h(\hat{\zeta}^{(u)})[1 + O(|\hat{\zeta} - \hat{\zeta}^{(u)}|)]}{h(\hat{\zeta}^{(u)})} = 1 + O\left(\frac{l}{u}\right). \tag{47}$$

To analyze the ratio of the second derivatives

$$\left[\frac{\dfrac{d^2}{d\zeta^2}\left.\left[\dfrac{1}{u}L(Q_u;\zeta) + \dfrac{l}{u}\dfrac{1}{l}L^{(1)}(Q_l;\zeta)\right]\right|_{\hat{\zeta}}}{\dfrac{d^2}{d\zeta^2}\left.L(Q_u;\zeta)\right|_{\hat{\zeta}^{(u)}}}\right]^{1/2}$$

recall that the third derivative of the average log-likelihood of the unlabeled samples is bounded near $\hat{\zeta}^{(u)}$. Consequently, from the intermediate value theorem it follows that

$$\begin{aligned}\frac{\dfrac{d^2}{d\zeta^2}\dfrac{1}{u}\left.L(Q_u;\zeta)\right|_{\hat{\zeta}}}{\dfrac{d^2}{d\zeta^2}\dfrac{1}{u}\left.L(Q_u;\zeta)\right|_{\hat{\zeta}^{(u)}}} &= [1 + O(\hat{\zeta} - \hat{\zeta}^{(u)}|)] \\ &= \left[1 + O\left(\frac{l}{u}\right)\right].\end{aligned}$$

Similarly

$$\frac{\dfrac{d^2}{d\zeta^2}\dfrac{l}{u}\dfrac{1}{l}\left.L^{(1)}(Q_l;\zeta)\right|_{\hat{\zeta}}}{\dfrac{d^2}{d\zeta^2}\dfrac{1}{u}\left.L(Q_u;\zeta)\right|_{\hat{\zeta}^{(u)}}} = \frac{l}{u}\,\frac{\dfrac{d^2}{d\zeta^2}\dfrac{1}{l}\left.L^{(1)}(Q_l;\zeta)\right|_{\hat{\zeta}}}{\dfrac{d^2}{d\zeta^2}\dfrac{1}{u}\left.L(Q_u;\zeta)\right|_{\hat{\zeta}^{(u)}}} = O\left(\frac{l}{u}\right).$$

Therefore, the ratio of the second derivatives in (45) is equal to

$$\begin{aligned}&\left[\frac{\dfrac{d^2}{d\zeta^2}\left.\left[\dfrac{1}{u}L(Q_u;\zeta) + \dfrac{l}{u}\dfrac{1}{l}L^{(1)}(Q_l;\zeta)\right]\right|_{\hat{\zeta}}}{\dfrac{d^2}{d\zeta^2}\left.L(Q_u;\zeta)\right|_{\hat{\zeta}^{(u)}}}\right]^{1/2} \\ &= \left[1 + O\left(\frac{l}{u}\right)\right]^{1/2} = \left[1 + O\left(\frac{l}{u}\right)\right].\end{aligned} \tag{48}$$

Consider now the quantity $\exp\{L(Q_u;\hat{\zeta}) - L(Q_u;\hat{\zeta}^{(u)})\}$. The second derivative of the average log-likelihood of the unlabeled samples in a neighborhood of $\hat{\zeta}^{(u)}$ is bounded from above. This can be checked easily with the same approach we have used for the third derivative. Since $\hat{\zeta}^{(u)}$ is a point of relative maximum of the log-likelihood

$$(1/u)(L(Q_u;\hat{\zeta}) - L(Q_u;\hat{\zeta}^{(u)})) = O(|\hat{\zeta}^{(u)} - \hat{\zeta}|^2) = O(l^2/u^2).$$

Thus

$$\exp\{L(Q_u;\hat{\zeta}) - L(Q_u;\hat{\zeta}^{(u)})\} = O\left(\frac{l^2}{u}\right). \tag{49}$$

Finally, consider the log-likelihood of the labeled samples. Recall that the derivative of $l^{-1}L^{(1)}(Q_l;\zeta)$ is conditionally bounded by a constant K in a neighborhood of $\hat{\zeta}^{(u)}$ given $|\hat{\zeta}^{(u)}-\zeta_0|<\delta$. Thus

$$\frac{1}{l}L^{(1)}(Q_l;\hat{\zeta}^{(u)})-M|\hat{\zeta}-\hat{\zeta}^{(u)}| \le \frac{1}{l}L^{(1)}(Q_l;\hat{\zeta}) \le \frac{1}{l}L^{(1)}(Q_l;\hat{\zeta}^{(u)})+M|\hat{\zeta}-\hat{\zeta}^{(u)}|$$

with probability converging exponentially fast to 1 in the number of unlabeled samples, and consequently

$$L^{(1)}(Q_l;\hat{\zeta})=L^{(1)}(Q_l;\hat{\zeta}^{(u)})+O\left(\frac{l^2}{u}\right). \tag{50}$$

The proof of (39) follows then immediately from (46)–(50). The same approach leads to the proof of (40) and to the last step in the derivation of the lemma. □

A4. Proof of Lemma 4

Lemma 4 states that risks of the optimal Bayesian classifier and of the classifier that uses the maximum-likelihood estimator $\hat{\zeta}^{(u)}$ of the parameter ζ based on the unlabeled samples alone are the same, up to terms of order $o(u^{-1})$.

We prove here the first part of the lemma, the proof of the second part being analogous. From the proof of Lemma 2 (Appendix A3) it follow that, as $l^{3+\epsilon}u^{-1}\to 0$

$$-\lim_{l,u\to\infty}\frac{1}{l}\log\Pr\left\{\frac{a}{\overline{a}}<2|Z=1\right\} = -\lim_{l,u\to\infty}\frac{1}{l}\log\Pr\left\{\frac{a}{\overline{a}}<1|Z=1\right\}.$$

Thus the training sets for which $\overline{a}\le a\le 2\overline{a}$ form a collection with probability

$$\Pr\{\overline{a}\le a\le 2\overline{a}|Z=1\}\le\Pr\{a<2\overline{a}|Z=1\} = \exp\{-lD+o(l)\}.$$

Their contribution to the probabilities of error of the optimal Bayesian classifier (6) and of test (19) is of the same order of term $\exp\{-lD+o(l)\}$ derived in Lemma 2. Thus the event $\{\overline{a}\le a<2\overline{a}\}$ will be accounted for in the expression (13) of the risk by the term $\exp\{-lD+o(l)\}$. Consider now all the training sets for which $a>2\overline{a}$. Let

$$t_1\triangleq E_h[\zeta|Q,Z=1]-\hat{\zeta}^{(u)}=b-\hat{\zeta}^{(u)}$$
$$t_2\triangleq E_h[\zeta|Q,Z=2]-\hat{\zeta}^{(u)}=c-\hat{\zeta}^{(u)}$$

where $b\triangleq E[\zeta|Q,Z=1]$, $c\triangleq E[\zeta|Q,Z=2]$, and $\hat{\zeta}^{(u)}$ is the maximum-likelihood estimator based on the unlabeled samples alone. Rewrite the ratio (14) used in the optimal Bayesian solution as

$$\frac{abg_1(X_0)+\overline{a}\overline{c}g_2(X_0)}{\overline{a}cg_1(X_0)+a\overline{b}g_2(X_0)} = \frac{a(\hat{\zeta}^{(u)}+t_1)g_1(X_0)+\overline{a}(\overline{\hat{\zeta}}^{(u)}-t_2)g_2(X_0)}{\overline{a}(\hat{\zeta}^{(u)}+t_2)g_1(X_0)+a(\overline{\hat{\zeta}}^{(u)}-t_1)g_2(X_0)}$$

and conclude that the boundaries of the decision regions of the Bayesian classifier (6) are the roots of the equation

$$a(\hat{\zeta}^{(u)}+t_1)g_1(X_0)+\overline{a}(\overline{\hat{\zeta}}^{(u)}-t_2)g_2(X_0) = \overline{a}(\hat{\zeta}^{(u)}+t_2)g_1(X_0)+a(\overline{\hat{\zeta}}^{(u)}-t_1)g_2(X_0)$$

or, equivalently, of

$$[a\hat{\zeta}^{(u)}+at_1-\overline{a}\hat{\zeta}^{(u)}-\overline{a}t_2]g_1(X_0) = [a\overline{\hat{\zeta}}^{(u)}-at_1-\overline{a}\overline{\hat{\zeta}}^{(u)}+\overline{a}t_2]g_2(X_0).$$

Therefore, the optimal Bayesian classifier (6) can be rewritten as

$$\begin{aligned}
\text{decide } \hat{\theta}_0=1 \quad \text{if } &(a-\overline{a})\left[\hat{\zeta}^{(u)}+\frac{at_1-\overline{a}t_2}{a-\overline{a}}\right]g_1(X_0)\\
&>(a-\overline{a})\left[1-\left(\hat{\zeta}^{(u)}+\frac{at_1-\overline{a}t_2}{a-\overline{a}}\right)\right]g_2(X_0)\\
\text{decide } \hat{\theta}_0=2 \quad \text{if } &(a-\overline{a})\left[\hat{\zeta}^{(u)}+\frac{at_1-\overline{a}t_2}{a-\overline{a}}\right]g_1(X_0)\\
&<(a-\overline{a})\left[1-\left(\hat{\zeta}^{(u)}+\frac{at_1-\overline{a}t_2}{a-\overline{a}}\right)\right]g_2(X_0).
\end{aligned} \tag{51}$$

By inspection one sees that test (16) based on the maximum-likelihood estimator and test (51) are structurally identical, the only difference being the substitution $\hat{\zeta}^{(u)}+(at_1-\overline{a}t_2)/(a-\overline{a})$ for $\hat{\zeta}^{(u)}$. This implies that both tests are structurally identical to test (9), the probability of error of which is shown in Theorem 1. Conditional on $Z=1$, $g_1(\cdot)=f_1(\cdot)$, $g_2(\cdot)=f_2(\cdot)$, and $\zeta=\eta$. Consequently, $\hat{\zeta}^{(u)}$ is actually an estimator of η, and we can write $\hat{\zeta}^{(u)}=\hat{\eta}^{(u)}$. Conditional on $\{a>2\overline{a}\}$ the denominator of $(at_1-\overline{a}t_2)/(a-\overline{a})$ is bounded away from 0, and the quantity is a *bona fide* random variable. We then can calculate the probability of error of test (51) by substituting

$$\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)}\triangleq\hat{\eta}^{(u)}+(at_1-\overline{a}t_2)/(a-\overline{a})$$

for $\hat{\eta}$ in the derivation of the probability of error of test (9). Using (10) we can write that, conditional on $\{a\ge 2\overline{a}\}$

$$\begin{aligned}
P_e(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)}) = {}&P_e(\eta_0)+\left.\frac{d}{d\hat{\eta}}P_e(\hat{\eta})\right|_{\eta_0}(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)}-\eta_0)\\
&+\left.\frac{d^2}{d\hat{\eta}^2}P_e(\hat{\eta})\right|_{\eta_0}\frac{(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)}-\eta_0)^2}{2}\\
&+r(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)})\frac{(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)}-\eta_0)^3}{6}.
\end{aligned}$$

Take expectations, recall that the first derivative of the probability of error is equal to zero, and conclude that

$$\begin{aligned}
R(l,u) = {}&R^*+\frac{1}{2}\left.\frac{d^2}{d\hat{\eta}^2}P_e(\hat{\eta})\right|_{\eta_0}E[(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)}-\eta_0)^2]\\
&+\frac{1}{6}r(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)})E[(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)}-\eta_0)^3].
\end{aligned} \tag{52}$$

Consider the first expectation. Expanding the square yields

$$E[(\hat{\eta}^{(u)}+\Delta\hat{\eta}^{(u)}-\eta_0)^2]=E[(\hat{\eta}^{(u)}-\eta_0)^2]+E[(\Delta\hat{\eta}^{(u)})^2] +2E[(\hat{\eta}^{(u)}-\eta_0)\Delta\hat{\eta}^{(u)}].$$

We can bound from above the last term by

$$|E[(\hat{\eta}^{(u)} - \eta_0)\Delta\hat{\eta}^{(u)}]| \leq \sqrt{E[(\hat{\eta}^{(u)} - \eta_0)^2]E[(\Delta\hat{\eta}^{(u)})^2]}$$

where

$$\sqrt{E[(\hat{\eta}^{(u)} - \eta_0)]} = O(u^{-1/2}).$$

From Lemma 5 it follows that $|t_1| \leq Ml/u$ and that $|t_2| \leq Ml/u$ with probability $\geq 1 - \exp\{-uD_\zeta\}$. Conditional on $\{a \geq 2\overline{a}\}$, it follows that the ratio

$$(at_1 - \overline{a}t_2)/(a - \overline{a}) \triangleq \Delta\hat{\eta}^{(u)}$$

is smaller in absolute value than $3M(l/u)$, again with probability $\geq 1 - \exp\{-uD_\zeta\}$. From the assumption that $l^{3+\epsilon}u^{-1} \rightarrow 0$ it follows that $\Delta\hat{\eta}^{(u)} = o(u^{-1/2})$ and that

$$E[(\hat{\eta}^{(u)} + \Delta\hat{\eta}^{(u)} - \eta_0)^2] = E[(\hat{\eta}^{(u)} - \eta_0)]^2 + o(u^{-1}).$$

A similar analysis holds for the remainder of the expansion of the risk, which allows us to conclude that the remainder is negligible. The derivation can be repeated conditioning on $Z_0 = 2$, to show that on the set $\{\overline{a} > 2a\}$ the additional risk $\Delta R(l, u)$ of the Bayesian solution (6) is asymptotically equivalent to the additional risk of test (19), namely, their ratio converges to one. □

ACKNOWLEDGMENT

The authors wish to thank the reviewers for reference [21] and for their useful comments.

REFERENCES

[1] V. Castelli and T. M. Cover, "The relative value of labeled and unlabeled samples in pattern recognition in the regular parametric case," in preparation, 1994.

[2] ———, "On the exponential value of labeled samples," *Patt. Recogn. Lett.*, vol. 16, pp. 105–111, Jan. 1995.

[3] ———, "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," Tech. Rep. 77, Stanford Univ., Dept. Statist., Mar. 1995.

[4] C. Chen, *Statistical Pattern Recognition.* New York: Hayden, 1973.

[5] M. F. da M. Tenorio, "Self-organizing neural network algorithm: Adapting structure for optimum supervised learning," in *Proc. Hawaii Int. Conf. on System Science*, 1990, vol. 1, pp. 187–193.

[6] A. Dembo and O. Zeitouni. *Large Deviations, Techniques and Applications.* Boston, MA: Jones and Bartlett, 1993.

[7] L. Devroye, "Automatic pattern recognition, a study of the probability of error," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 10, no. 4, pp. 530–543, July 1988.

[8] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[9] K. Fukanaga. *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 1972.

[10] H. Greenspan, R. Goodman, and R. Chellappa, "Texture analysis via unsupervised and supervised learning," in *Int. Joint Conf. on Neural Networks IJCNN 91* (Seattle, WA, July 8–12, 1991), vol. 1, pp. 639–644.

[11] R. M. Holdaway, "Enhancing supervised learning algorithms via self-organization," in *Int. Joint Conf. on Neural Networks IJCNN 89* (June 18-22 1989), pp. 523–530.

[12] A. K. Jain and B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice," in P. R. Krishnaian and L. N. Kanal, Eds., *Handbook of Statistics*, vol. 2. Amsterdam, The Netherlands: North-Holland, 1982, pp. 1464–1469.

[13] R. E. Kass, L. Tierney, and J. B. Kadane, "The validity of posterior expansions based on Laplace's method," in *Bayesian and Likelihood Methods in Statistics and Econometrics*, 1990, pp. 473–488.

[14] ———, "Laplace's method in Bayesian analysis," *Contemp. Math.*, vol. 115, pp. 89–99, 1991.

[15] E. L. Lehmann, *Theory of Point Estimation.* New York: Wiley, 1983.

[16] ———, *Testing Statistical Hypotheses.* Wadsworth & Brooks/Cole, 1991.

[17] G. J. McLachlan, "Estimating the linear discriminant function from initial samples containing a small number of unclassified observations," *J. Amer. Statist. Assoc.*, vol. 72, no. 358, pp. 403–406, June 1977.

[18] G. J. McLachlan and S. Ganesalingam, "Updating the disctiminant function on the basis of unclassified data," *Commun. Statist.-Simul.*, vol. 11, no. 6, pp. 753–767, 1982.

[19] J. M. Mendel and K. S. Fu, *Adaptive, Learning, and Pattern Recognition Systems: Theory and Applications.* New York: Academic Press, 1970.

[20] J. M. J. Murre, R. H. Phaf, and G. Wolters. "CALM networks: A modular approach to supervised and unsupervised learning," in *Int. Joint Conf. on Neural Networks IJCNN 89* (June 18-22, 1989), pp. 649–656.

[21] T. O'Neill, "The general distribution of the error rate of a classification procedure with application to logistic regression discrimination," *J. Amer. Statist. Assoc.*, vol. 75, no. 379, pp. 154–169, Mar. 1980.

[22] ———, "Normal discrimination with unclassified observations," *J. Amer. Statist. Assoc.*, vol. 73, no. 364, pp. 821–826, Dec. 1978.

[23] Y.-H. Pao and D. J. Sobajic, "Combined use of supervised and unsupervised learning for dynamic security assessment," *IEEE Trans. Power Syst.*, vol. 7, no. 2, pp. 878–884, May 1992.

[24] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Reccomendations for practitioners," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 13, no. 3, pp. 252–268, Mar. 1991.

[25] A. SenGupta, "A review of optimality of multivariate tests," *Statist. Probab. Lett.*, vol. 12, pp. 527–535, Dec. 1991.

[26] B. M. Shahshahani and D. A. Landgrebe, "On the asymptotic improvement of supervised learning by utilizing additional unlabeled samples; normal mixture density case," in *Neural and Stochastic Methods in Image Signal Processin, Proc. SPIE* (July 20–23, 1992), vol. 1766, pp. 143–155.

[27] C. W. Therrien, *Decision, Estimation and Classification.* New York: Wiley, 1989.

[28] V. V. Tolat and A. M. Peterson, "Nonlinear mapping with minimal supervised learning," in *Proc. Hawaii Int. Conf. on System Science* (Jan. 2–5, 1990), pp. 170–179.

[29] C. W. Therrien, *Decision, Estimation, and Classification: An Introduction to Pattern Recognition and Related Topics.* New York: Wiley, 1989.