

Minimum Probability of Error Signal Representation based on a Rate-Distortion Optimality Criterion

Jorge Silva and Shrikanth Narayanan, *Senior Member, IEEE*

Department of Electrical Engineering, Signal and Image Processing Institute

University of Southern California, Viterbi School of Engineering

3740 McClintock Avenue, Room EEB430

Los Angeles, CA 90089 2564

Email: *jorgesil@usc.edu*, *shri@sipi.usc.edu*

Web: <http://sail.usc.edu/>

Tel: 213-740-3477, Fax: 213-740-4651

Abstract

The focus of this work is to study the role of signal representation in pattern recognition from an information theoretic perspective. This work addresses the problem of minimum probability of error signal representation (MPE-SR) considering issues of finite training data. Theoretical results originally presented by Vasconcelos [1] are extended in good generality that justifies addressing the MPE-SR as a complexity-regularized optimization problem, reflecting the well known tradeoff between signal representation quality and learning complexity. A rate-distortion formulation is proposed to address this optimization problem by finding a sequence of signal representations achieving optimal complexity-fidelity operational points. Finally, it is formally shown that under specific assumptions the MPE-SR principle reduces to two emblematic scenarios well-known in pattern recognition: pruning algorithms for classification trees (CART) [2], and some versions of Fisher linear discriminant analysis.

Index Terms

Signal representation, feature extraction, pattern recognition, Bayes decision approach, complexity regularization, rate-distortion theory, statistical learning theory, decision trees, linear discriminant analysis, mutual information.

I. INTRODUCTION

The notion of optimal signal representation is a fundamental problem that the signal processing community has been addressing from different angles and in multiple research context. The formulation and solution of this problem have provided significant contributions in lossy compression, estimation and denoising problems [3]–[7]. The overarching motivation is to find parsimonious representations for a particular family of signals, where the fact that few coordinates capture most of the target signal energy has proved to be of significant importance in

improving compression and denoising techniques [6]–[8]. In this regard, defining an optimal basis that represents the nature of a given source or even more, dynamically selecting bases that adapt to the statistics of non-stationary sources — non-linear approximation— have been topics of significant research in the signal processing community [3], [8], [9]. Emblematic representation bases include the wavelet families [10], discrete cosine transform (DCT), and more general over complete families, particularly those useful in the context of non-linear approximation [3]. Over-complete representation introduces the issue of redundancy — more representation elements in the dictionary than the dimension of the signal space — and consequently the problem of finding an accurate sparse representation for a given signal is a difficult task and a topic of current research interest [11]–[13].

In the context of pattern recognition, signal representation issues are naturally associated with feature extraction (FE). In contrast to compression and denoising scenarios, where the objective is to design bases that allow optimal representation of the actual observation source using fidelity metrics like the mean square error, in recognition we are looking for representations that capture an unobserved phenomena that need to be inferred from the observed signal. Consequently, the optimality criterion is associated with minimizing the probability of error of taking the mentioned decision or generally minimizing an average risk indicator, for example, using a Bayes decision approach [2], [14]–[16]. In this context, the observation signal can be considered a combination of multiple sources of innovation, not all of them related with the underlying target phenomenon. Hence, from a signal representation point of view, one objective is to characterize the observation subspace which is relevant for the decision problem, the well known concept of sufficient statistics for the Bayes decision theoretic approach [17]–[19]. An example of the use of sufficient statistics for feature representation is the basic detection problem formulated in communication theory [17]. In this scenario, the signal constellation and the statistics of the channel are known, and consequently there is an analytical solution for the observation sub-space which provides sufficient statistics for the detection problem, the well-known matching filter [17]. In pattern recognition, we are dealing with a more complex scenario, because we typically do not know the generative process in which different sources of information are combined to generate the observation phenomena and we need to address the problem of feature extraction in an unsupervised way.

In classification problems, the joint distribution between the observation and the underlying target phenomenon is generally not available, and it needs to be estimated based on a finite amount of training data [2], [14], [15], [20]. It is well known that the accuracy of this estimation process is affected by the dimensionality of the observation space — the curse of dimensionality — which is proportional to the disagreement between the real and the estimated distributions for a given finite amount of trained data. In this particular regard, Vasconcelos [1] has recently formalized the effect of estimation error by an information theoretic quantity, a function of the Kullback-Leibler divergence (KLD) [18] between the class-conditional probabilities associated with the empirical and real distributions. Then, an integral part of the feature extraction (FE) phase is to address the problem of optimal dimensionally reduction, particularly necessary in scenarios where the original raw observation measurements lie in a high-dimensional space, and typically only a limited amount of training data are available, such as in most speech classification [21], image classification [22] and hyper-spectral classification scenarios [23], [24]. Consequently, finding signal representations that allow capturing the target phenomenon in a relatively low-dimensional description

is a crucial aspect of the problem. This is conceptually related with the problem of finding parsimonious signal representations, but with an objective criterion which is clearly different from the classical distortion measures used in lossy compression [25]–[27]. The goal here is signal classification rather than signal reproduction.

The feature extraction (FE) problem in most of the cases is based on particular knowledge of the task. This knowledge is used to characterize potentially salient features. For example, in the case of speech recognition, a short-term spectral envelope of the speech provides useful phonetic discrimination information [28]. However, there are some problems in which it is not possible to characterize the set of relevant features in advance. An optimality principle that can be used to select those salient features from a relatively large collection of potential representations, hence is an interesting and relevant problem in pattern recognition. Many algorithms have been proposed along this direction for finding feature transformations that minimize some optimality criterion directly or indirectly associated with the probability of error, for instance, information measures like the KLD or the divergence [29] and mutual information [21], empirical discriminant measures like the Mahalanobis distance and Fisher’s class separability metric [14], [22], and the empirical Bayes error using cross validation [1], [2]. Most of those algorithms address the problem of feature selection (FS), where the idea is to select a sub-set of the more informative features from a given feature collection. However even in this simple case, only sub-optimal greedy algorithms have been proposed [30], because finding a solution for the FS problem implies an exhaustive combinatorial search [31], which is prohibitive for many real scenarios. Proposed solutions for the FE problem impose assumptions on the family of feature transformations — parametric or non-parametric approaches for estimating the joint class-observation distributions — and the optimality criterion, which allow to approximate or find closed-form solutions in a particular application domain, where the modeling assumptions can be considered valid.

Despite issues in finding feature representations of lower complexity which capture the most discriminant aspects of the full measurement-observation space, the problem is a well motivated one and good approximations have been presented under certain modeling assumptions [1], [21], [29], [30]. Nevertheless, there has not been a concrete general formulation of the ultimate problem, which is to find the minimum probability error signal representation (MPE-SR) constrained on a given amount of training data or any additional operational cost that may constrain the classification task. Such a formulation would provide a better theoretical support and justification for the aforementioned FE problem in pattern recognition under different operational scenarios. New and concrete results in the direction of formalizing the MPE-SR problem have been recently presented by Vasconcelos [1]. [1] formalizes a tradeoff between the Bayes error bound and an information-theoretic indicator of the estimation error because of the use of finite empirical data, and connects this result with the concept of optimal signal representation for classification.

The present work was motivated and built upon the original ideas presented in [1]. The paper addresses the MPE-SR using a rate-distortion formulation, by drawing a natural analogy of this problem to lossy compression with a fidelity criterion [19], [25]–[27], [32]. We start by extending results for the MPE-SR problem presented by Vasconcelos [1]. The formulation of the problem considers a set of representations characterized by a rich family of transformations of the original raw observation space, where the objective is to find the one that minimizes the

probability of error with a given amount of training data. For doing that, we extend in more generality the result presented by Vasconcelos [1], that shows a tradeoff between Bayes error and estimation error across a sequence of feature transformations of increasing complexity — measured in terms of dimensionality or cardinality of the transformed space depending on the context — with a strong embedded structure. In particular, the present work extends this result for a general family of feature-embedded representations and formalizes sufficient conditions for those results to hold, which not only takes into consideration the embedded structure of the feature representation family, as originally considered in [1], but also the consistent nature of the family of empirical class-observation distributions estimated across the sequence of observation representations.

The Bayes estimation error tradeoff is used to formulate the problem of MPE-SR as a complexity-regularized optimization problem, with an objective function that considers a fidelity indicator, which represents the Bayes error, and a cost or penalization indicator — associated with the complexity of the representation — which reflects the estimation error. We show that the solution of this problem relies on a particular sequence of representations, which is the solution of an operational rate-distortion problem. Interestingly, we show that the well known CART pruning algorithm [2] and Fisher linear discriminant analysis [14], are particular instances of this rate-distortion formulation, and we formally justify this connection.

The rest of the paper is organized as follows. Section II presents the general problem formulation, terminologies and review of some results that will be used in this work. Section III introduces the main theoretical results presented in this work; in particular it shows the important tradeoff between Bayes error and estimation error across a sequence of feature spaces of increasing complexity. Section IV presents the rate-distortion formulation and Section V, a formal connection of this problem with the MPE-SR problem. Sections VI and VII show how the MPE-SR principle can be particularized into a classification tree problem (CART learning algorithms) and linear discriminant analysis, respectively. Section VIII offers final discussion and a conceptual connection of the MPE-SR with the *empirical risk minimization* (ERM) and *structural risk minimization* (SRM) principle [15], [16], [20]. Finally, Section IX provides future directions.

II. PRELIMINARIES: BAYES DECISION APPROACH

Let $X(u):(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ be a random vector taking values in a finite dimensional Euclidian space $\mathcal{X} = \mathbb{R}^K$, for some $K \in \mathbb{N}$, and $Y(u):(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ be a random variable taking values in a finite alphabet space \mathcal{Y} ¹. $(\Omega, \mathcal{F}, \mathbb{P})$ is the underlying probability space. We refer to $X(u)$ and $Y(u)$ as the observations and the class label random phenomena, respectively. The joint observation-class random vector $(X(u), Y(u))$ induces a joint probability measure $P_{X,Y}$ in the measurable space $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_{\mathcal{X}} \times \mathcal{F}_{\mathcal{Y}}))$ ². Knowing the joint distribution $P_{X,Y}$, the problem is to find a measurable decision function $g(\cdot)$ from $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ to $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ such that given realizations of $X(u)$, infer their discrete counterpart $Y(u)$ with the minimum expected cost, or minimum risk, given by $\mathbb{E}_{X,Y} [l(g(X), Y)]$.

¹In this particular context, given that $\mathcal{X} = \mathbb{R}^K$, a natural choice for $\mathcal{F}_{\mathcal{X}}$ is the Borel sigma field [34], and for $\mathcal{F}_{\mathcal{Y}}$ the power set of \mathcal{Y} .

² $\sigma(\mathcal{F}_{\mathcal{X}} \times \mathcal{F}_{\mathcal{Y}})$ refers to the product sigma field that makes the coordinate projection on the labeling and observation spaces, $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ respectively, measurable [34].

In this context, $l(y_1, y_2)$ represents the penalization of labeling an observation with the value y_1 , when its true label is given by y_2 , $\forall y_1, y_2 \in \mathcal{Y}$. The minimum risk decision is called the Bayes decision rule, where for the emblematic 0-1 delta risk function, $l(y_1, y_2) = \delta(y_1, y_2)$, the Bayes rule minimizes the probability error that is equivalent to the maximum a posteriori decision (MAP) rule, Eq.(1).

$$g_{P_{X,Y}}(\bar{x}) \equiv \arg \max_{y \in \mathcal{Y}} P_{X,Y}(\bar{x}, y), \quad \forall \bar{x} \in \mathcal{X} \quad (1)$$

In this case the optimal error probability, $L_{\mathcal{X}} \equiv \mathbb{P}(\{u \in \Omega : g_{P_{X,Y}}(X(u)) \neq Y(u)\})$, is given by [35]

$$\begin{aligned} L_{\mathcal{X}} &= P_{X,Y}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : g_{P_{X,Y}}(x) \neq y\}) \\ &= 1 - \mathbb{E}_{X,Y}(\mathbb{1}_{\{(x,y) \in \mathcal{X} \times \mathcal{Y} : g_{P_{X,Y}}(x) = y\}}(X, Y)) \\ &= 1 - \mathbb{E}_X \left(\mathbb{E}_{Y|X} \left(\mathbb{1}_{\{(x,y) \in \mathcal{X} \times \mathcal{Y} : g_{P_{X,Y}}(x) = y\}}(X, Y) | X \right) \right) \\ &= 1 - \mathbb{E}_X \left[\max_{i \in \mathcal{Y}} P_{Y|X}(i|X) \right], \end{aligned} \quad (2)$$

where $\mathbb{1}_A(\cdot)$ is the indicator function.

This is the optimal performance that we can get as a function of the discrimination representation of the observation space \mathcal{X} , that Vasconcelos [1] denoted as the Bayes error bound. The subscript notation on $L_{\mathcal{X}}$ represents the fact that this optimal probability of error is a function of the representation quality of \mathcal{X} . Note that the Bayes decision rule $g_{P_{X,Y}}(\cdot)$ [2], induces an optimal partition in the observation space \mathcal{X} , given by $\mathcal{V}_{\mathcal{X}} \equiv \{g_{P_{X,Y}}^{-1}(\{y\}) : y \in \mathcal{Y}\}$. The following lemma states a version of the well known result that any deterministic measurable transformation of the observation space \mathcal{X} can not provide discrimination gain — in the sense of improving the probability of error — under the assumption that we know the class-observation distribution $P_{X,Y}$.

LEMMA 1: (*Theorem 3*, [1]) Let $(X(u), Y(u))$ be a random vector with joint distribution $P_{X,Y}$ defined on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_{\mathcal{X}} \times \mathcal{F}_{\mathcal{Y}}))$. Consider $\mathbb{f} : (\mathcal{X}, \mathcal{F}_{\mathcal{X}}) \rightarrow (\mathcal{X}', \mathcal{F}'_{\mathcal{X}})$ to be a measurable mapping. If we define $X'(u) \equiv \mathbb{f}(X(u))$ as a new observation random variable, with joint probability distribution $P_{X',Y}$ induced by $\mathbb{f}(\cdot)$ and the original probability space $((\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_{\mathcal{X}} \times \mathcal{F}_{\mathcal{Y}}), P_{X,Y})$, we have that

$$L_{X'} = \mathbb{P} \left(\left\{ u \in \Omega : g_{P_{X',Y}}(X'(u)) \neq Y(u) \right\} \right) \geq L_X. \quad (3)$$

A version of the proof of this well known result was presented in [36].

Based on this result, it is natural to state the concept of sufficient statistics in the context of measurable transformation of the observation space.

Definition 1: A measurable transformation $\mathbb{f}(\cdot) : (\mathcal{X}, \mathcal{F}_{\mathcal{X}}) \rightarrow (\mathcal{X}', \mathcal{F}'_{\mathcal{X}})$ provides sufficient statistics for inferring the class random variable $Y(u)$ if, and only if,

$$L_{X'} = L_X.$$

Remark 1: The Bayes classification rule in the transform domain, $g_{P_{X',Y}}(\cdot)$, induces a measurable partition on the original observation space \mathcal{X} given by $\mathcal{V}_{\mathcal{X}'} \equiv \{\mathbb{f}^{-1}(g_{P_{X',Y}}^{-1}(\{y\})) : y \in \mathcal{Y}\}$. Then, the performance degradation can be seen as the statistical mis-match between $\mathcal{V}_{\mathcal{X}'}$ and $\mathcal{V}_{\mathcal{X}}$. In particular, it is straightforward to prove

that a sufficient condition for $\mathbb{f}(\cdot)$ being sufficient statistics is that $\mathcal{V}_{\mathcal{X}'}$ is equal to $\mathcal{V}_{\mathcal{X}}$ P_X -almost surely, i.e., $P_X \left(\bigcup_{y \in \mathcal{Y}} \left[\mathbb{f}^{-1}(g_{P_{X',Y}}^{-1}(\{y\})) \triangle g_{P_{X,Y}}^{-1}(\{y\}) \right] \right) = 0$.

In real scenarios we do not have access to the true joint distribution $P_{X,Y}$, but instead we have iid realizations of $(X(u), Y(u))$, $D_N \equiv \{(x_i, y_i) : i \in \{1, \dots, N\}\}$, which in the Bayes approach are used to characterize an estimation of the joint observation-class distribution, the empirical distribution denoted by $\hat{P}_{X,Y}$. This estimated distribution $\hat{P}_{X,Y}$ is used to define an empirical Bayes classification rule, by Eq.(1), that we denote as $\hat{g}_{\hat{P}_{X,Y}}(\cdot)$. Note that the average risk of the empirical Bayes rule, Eq.(4), differs from the optimal Bayes error bound $L_{\mathcal{X}}$ as a consequence of the estimation error.

$$\mathbb{P} \left(\left\{ u \in \Omega : \hat{g}_{\hat{P}_{X,Y}}(X(u)) \neq Y(u) \right\} \right) \quad (4)$$

It is well understood that the estimation error introduces performance degradation with respect to the Bayes error bound $L_{\mathcal{X}}$, and that the magnitude of this error is a function of some notion of complexity of the observation space constrained to a finite amount of training data [1], [23], [37]. This mostly explains the well known “peaking or Hughes phenomenon” [38], in which for a given amount of training data, D_N , by increasing the complexity of the observation space, the system performance increases until a certain point where it starts presenting systematic degradation — overfitting effects [14], [30], [38]–[40]. This implies a strong relationship between the number of training examples and a notion of complexity of the observation space, as well known in the justification of dimensionality reduction as a fundamental part of feature extraction. This inter-relationship between observation space complexity and the amount of training data is also a function of the modeling assumptions made in learning the joint observation-class distribution.

In this work, we mainly focus on studying aspects of optimal feature representation for classification, assuming the presented Bayes decision approach and that the learning framework satisfies some desirable conditions that will be specified in the next section. Under this assumption, we can consider two underlying factors associated with signal representation of the observation space that affect the performance of a Bayes decision framework. One relates to the signal representation quality associated with the Bayes error bound, and the other, the signal space complexity associated with the estimation error for given finite training data [1]. As a consequence, it is natural to think that, having a rich family of feature representations of the raw observation phenomenon, there is one in which this tradeoff between “representation quality” and “complexity” achieves an optimal solution for the error probability or average risk of the empirical Bayes classification rule, Eq.(4). The formal formulation of this tradeoff is the topic of the following sections and the main conceptual focus of this work. In this direction, the next section summarizes and extends some of the results originally presented by Vasconcelos [1] in the context of finding *minimum probability of error signal representation* (MPE-SR) for the problem of content-based image retrieval using a Bayes decision approach.

III. RESULT ON SIGNAL REPRESENTATION FOR CLASSIFICATION

We extend two important results associated with the role of signal representation for classification presented in [1]. Those results provide a clear understanding of the implicit role of the Bayes error bound and the estimation

errors for the presented Bayes decision approach. More importantly in the context of this exposition, it is shown how varying the representation complexity in a sequence of embedded spaces produces a tradeoff between the Bayes error bound and estimation error.

A. KLD as information quantity for the estimation error in a Bayes decision approach

The following theorem originally presented in [1] characterizes an information theoretic indicator of the estimation error, which upperbounds the performance degradation with respect to the Bayes error bound.

THEOREM 1: (*Theorem 4*, [1]) Let us consider a joint observation-class distribution $P_{X,Y}$ and an empirical joint distribution $\hat{P}_{X,Y}$, both defined on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_X \times \mathcal{F}_Y))$, which only differs in their class conditional probabilities (i.e., $\hat{P}_Y(\{y\}) = P_Y(\{y\})$, $\forall y \in \mathcal{Y}$). Then, the following inequality holds involving the performance of the empirical MAP decision rule $\hat{g}(\cdot)$, the optimal Bayes bound, Eq.(2), and an information theoretic indicator of the estimation error:

$$\mathbb{P} \left(\left\{ u \in \Omega : \hat{g}_{\hat{P}_{X,Y}}(X(u)) \neq Y(u) \right\} \right) - L_{\mathcal{X}} \leq \Delta g_{MAP}(\hat{P}_{X,Y}) \quad (5)$$

where

$$\Delta g_{MAP}(\hat{P}_{X,Y}) \equiv \sqrt{2 \ln 2} \sum_{y \in \mathcal{Y}} P_Y(\{y\}) \cdot \sqrt{\min \left\{ D(P_{X|Y}(\cdot|y) || \hat{P}_{X|Y}(\cdot|y)), D(\hat{P}_{X|Y}(\cdot|y) || P_{X|Y}(\cdot|y)) \right\}} \quad (6)$$

and $D(\cdot || \cdot)$ is the Kullback-Leibler divergence (KLD) [18] between two probability distributions on $(\mathcal{X}, \mathcal{F}_X)$, given by³,

$$D(P^1 || P^2) = \int_{\mathcal{X}} p^1(x) \cdot \log \frac{p^1(x)}{p^2(x)} \partial x$$

where p^1 and p^2 are the pdfs of P^1 and P^2 , respectively ⁴.

Vasconcelos proves this result for the case when the classes are equally likely [1] (*Theorem 4*). *Appendix I* presents the proof for the general case stated in **THEOREM 1**.

$\Delta g_{MAP}(\hat{P}_{X,Y})$ is the P_Y -average of a non-decreasing function of the Kullback-Leibler divergences (KLD) between the conditional class probabilities and their empirical counterpart. The KLD has a well known interpretation as a statistical discrimination measure between two probabilistic models [18], [19], [41], however in this case, it is an indicator of the performance deviation, relative to the fundamental performance bound, as a consequence of the statistical mismatch occurring in estimating the class-conditional probabilities.

Remark 2: $\Delta g_{MAP}(\hat{P}_{X,Y})$ is well defined if the conditional class distributions are absolute continuous with respect to the other associated distributions, Eq.(6). This assumption is not unreasonable in this scenario because the empirical joint distribution is induced by iid realizations of the true distribution. This condition is assumed for the rest of the exposition making all the KLD terms in Eq.(6) well defined.

³A necessary and sufficient condition for the KLD of P_X^1 with respect to P_X^2 to be well defined, is that $P_X^1 \ll P_X^2$ [18].

⁴We assume that the involved distribution are absolutely continuous with respect to the Lebesgue measure for defining the KLD in this domain [41].

The next result shows the natural extension of THEOREM 1 for the important case when the observation random variable $X(u)$ takes values in a finite alphabet space, denoted by \mathcal{A}_X . This result will be relevant when considering the case of a family of finite alphabet feature spaces induced by quantizing a raw continuous observation space, classification trees being an instance of this case [2].

COROLLARY 1: Let $(X(u), Y(u))$ be a random vector taking values in the finite product space $\mathcal{A}_X \times \mathcal{Y}$, with $P_{X,Y}$ and $\hat{P}_{X,Y}$ being the probability and the empirical probability, respectively. Assuming that $P_{X,Y}$ and $\hat{P}_{X,Y}$ only differ in their class-conditional probabilities, then Eqs. (5) and (6) hold, yielding $D(P_{X|Y}(\cdot|y) || \hat{P}_{X|Y}(\cdot|y))$ the discrete version of the KLD to be given by [18], [41]:

$$D(P_{X|Y}(\cdot|y) || \hat{P}_{X|Y}(\cdot|y)) = \sum_{x \in \mathcal{A}_X} P_{X|Y}(x|y) \cdot \left(\log \frac{P_{X|Y}(x|y)}{\hat{P}_{X|Y}(x|y)} \right).$$

The proof of this result can be directly derived from the one presented in *Appendix I*.

B. Tradeoff between Bayes error bound and the Estimation error

The following result connects aspects of signal representation into the classification problem by characterizing a tradeoff between the Bayes error bound and the estimation error. Before that, it is important to introduce the notion of an embedded space sequence, which provides a sort of order relationship among a family of feature observation spaces. Also we introduce the notion of consistent probability measures associated with an embedded space structure.

Definition 2: Let us consider a family of measurable transformations $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ from the same domain, $(\mathcal{X}, \mathcal{F}_\mathcal{X})$, but taking values in $\{\mathcal{X}_1, \dots, \mathcal{X}_n\}$, sequence of spaces of strictly increasing finite dimensionality, i.e., $\dim(\mathcal{X}_i) < \dim(\mathcal{X}_{i+1}), \forall i \in \{1, \dots, n-1\}$. The family of transformations $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ is called dimensional embedded if $\forall i \in \{1, \dots, n-1\}, \exists \pi_{i+1,i}(\cdot)$ measurable mapping from $(\mathcal{X}_{i+1}, \mathcal{F}_{i+1})$ to $(\mathcal{X}_i, \mathcal{F}_i)$ ⁵ such that,

$$\mathbb{F}_i(x) = \pi_{i+1,i}(\mathbb{F}_{i+1}(x)), \quad \forall x \in \mathcal{X}.$$

In this context, we also say that $\{\mathcal{X}_1, \dots, \mathcal{X}_n\}$ is dimensional embedded with respect to $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ and $\{\pi_{i+1,i}(\cdot) : i = 1, \dots, n-1\}$. Those dependencies will be implicit when not mentioned.

Definition 3: Let us consider $\{\mathcal{X}_i : i = 1, \dots, n\}$ to be a sequence of dimensional embedded spaces, where $\pi_{i+1,i} : (\mathcal{X}_{i+1}, \mathcal{F}_{i+1}) \rightarrow (\mathcal{X}_i, \mathcal{F}_i)$ is the measurable mapping stated in *Definition 2*. Associated with those spaces, let us consider a probability measure \hat{P}_i defined on $(\mathcal{X}_i, \mathcal{F}_i), \forall i \in \{1, \dots, L\}$. The family of probability measures $\{\hat{P}_i : i = 1, \dots, n\}$ is consistent with respect to the embedded sequence if

$$\forall i, j \in \{1, \dots, n\}, i < j, \forall B \in \mathcal{F}_i$$

$$\hat{P}_i(B) = \hat{P}_j(\pi_{j,i}^{-1}(B))$$

where $\pi_{j,i}(\cdot) \equiv \pi_{j,j-1}(\pi_{j-1,j-2}(\dots \pi_{i+1,i}(\cdot) \dots))$.

⁵For all practical purposes \mathcal{X}_i is a finite dimensional Euclidian space and \mathcal{F}_i refers to the Borel sigma filed.

Definition 3 is equivalent to saying that if we induce a probability measure on $(\mathcal{X}_i, \mathcal{F}_i)$ by using the measurable mapping $\pi_{j,i}(\cdot)$ and the probability measure \hat{P}_j on the space $(\mathcal{X}_j, \mathcal{F}_j)$, the induced measure is equivalent to \hat{P}_i . Consequently, the probabilistic description of the sequence of embedded spaces is univocally characterized by the more informative probability space, $(\mathcal{X}_n, \mathcal{F}_n, \hat{P}_n)$, and the family of measurable mappings $\{\pi_{j,i}(\cdot) : j > i\}$ of the embedded structure presented in *Definition 2*.

THEOREM 2: Consider $(X(u), Y(u))$ to be the joint class-observation random vector with joint distribution $P_{X,Y}$ defined on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_X \times \mathcal{F}_Y))$, where \mathcal{X} is a finite dimensional space \mathbb{R}^K and $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ a sequence of representation functions, with $\mathbb{F}_i(\cdot) : (\mathcal{X}, \mathcal{F}_X) \rightarrow (\mathcal{X}_i, \mathcal{F}_i)$, measurable $\forall i \in \{1, \dots, n\}$. In addition, let us assume that, $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ is a family of dimensional embedded transformations, satisfying $\mathbb{F}_i(\cdot) = \pi_{j,i}(\mathbb{F}_j(\cdot))$ for all $j > i$ in $\{1, \dots, n\}$. Then, considering the family of representation random variables $\{X_i(u) = \mathbb{F}_i(X(u)) : i = 1, \dots, n\}$ the Bayes error bound satisfies the following relationship across the sequence of embedded spaces:

$$L_{\mathcal{X}_{i+1}} \leq L_{\mathcal{X}_i}, \quad (7)$$

$\forall i \in \{1, \dots, n-1\}$.

If in addition we have a family of empirical probability measures $\{\hat{P}_{X_i, Y} : i = 1, \dots, n\}$, where $\hat{P}_{X_i, Y}$ is a probability measure on the space $(\mathcal{X}_i \times \mathcal{Y}, \sigma(\mathcal{F}_i \times \mathcal{F}_Y))$, with conditional class distribution families $\{\hat{P}_{X_i|Y}(\cdot|y) : i = 1, \dots, n\}$ consistent with respect to the embedded space sequence $\{\mathcal{X}_i : i = 1, \dots, n\} \forall y \in \mathcal{Y}$, then the following relationship for the estimation error applies:

$$\Delta g_{MAP}(\hat{P}_{X_i, Y}) \leq \Delta g_{MAP}(\hat{P}_{X_{i+1}, Y}), \quad (8)$$

$\forall i \in \{1, \dots, n-1\}$.

This result presents a tradeoff between the Bayes bound and estimation errors by considering a family of representations of monotonically increasing complexity. In other words, by increasing complexity we improve the theoretical performance bound — Bayes error bound — that we can potentially achieve, but as a consequence of increasing the estimation error, which upper bounds how far we can potentially be from the optimal Bayes bound, per *Theorem 1*. Given that the sequence is embedded, the Bayes error inequality, Eq.(7), is a direct consequence that lower dimensional representation rvs. are deterministic functions of the higher dimensional ones. For the estimation inequality, the main assumption is that the empirical distribution induced by the training data satisfies the defined consistency condition, and under this assumption the estimation error is proportional to the dimensionality of the feature space. The proof of this result is presented in *Appendix II*.

The following corollary states the original result presented in [1] (*Theorem 5*) for the tradeoff between Bayes error and estimation errors. This was presented in the important case when the embedded feature spaces are induced by coordinate projections. In this scenario, the consistency condition of the empirical distribution is natural for almost any probability estimation technique and, consequently, this condition is assumed and not included in the statement as originally presented in [1].

COROLLARY 2: (*Theorem 5*, [1]) Let $\mathcal{X} = \mathbb{R}^K$ and the family of coordinate projection $\pi_m^K(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^m$, $m \leq K$, be given by: $\pi_m^K(x_1, \dots, x_m, \dots, x_K) = (x_1, \dots, x_m)$, $\forall (x_1, \dots, x_K) \in \mathbb{R}^K$. Let $P_{X,Y}$ and $\hat{P}_{X,Y}$ be the joint probability measure and its empirical counterpart, respectively, defined on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_X \times \mathcal{F}_Y))$. Given that the coordinate projections are measurable, it is possible to induce those distributions on the sequence of embedded subspaces $\{\mathcal{X}_1, \dots, \mathcal{X}_K\}$ characterized by: $\mathcal{X}_i = \pi_i^K(\mathcal{X})$, $\forall i \in \{1, \dots, K\}$. Then, across this sequence of dimensional embedded spaces, the Bayes bound and estimation error satisfy the following inequalities, respectively:

$$L_{\mathcal{X}_{i+1}} \leq L_{\mathcal{X}_i}$$

$$\Delta g_{MAP}(\hat{P}_{\mathcal{X}_{i+1}, Y}) \geq \Delta g_{MAP}(\hat{P}_{\mathcal{X}_i, Y})$$

$\forall i \in \{1, \dots, K-1\}$.

Remark 3: From this corollary a natural approach to ensure that the family of empirical class conditional distributions $\{\hat{P}_{\mathcal{X}_i|Y}(\cdot|y) : i = 1, \dots, n\}$ are consistent across a dimensional embedded space sequence $\{\mathcal{X}_i : i = 1, \dots, n\}$, where the Bayes-estimation error tradeoff is manifested per THEOREM 2, is to constructively induce $\hat{P}_{\mathcal{X}_i|Y}(\cdot|y)$ using the empirical distribution on the most informative representation space, $\hat{P}_{\mathcal{X}_n|Y}(\cdot|y)$ on $(\mathcal{X}_n, \mathcal{F}_n)$, and the measurable mappings $\pi_{n,i}(\cdot)$, $\forall i < n$ associated with the embedded space sequence, *Definition 2*. This construction is appealing in particular when assuming parametric class conditional distributions like Gaussian mixture models (GMMs) and simple family of transformations like linear transformations, where inducing those distributions implies simpler operation on parameters of $\hat{P}_{\mathcal{X}_n|Y}(\cdot|y)$. This scenario was implicitly considered in [1].

As in the case of THEOREM 1, we also extend THEOREM 2 for the case when the family of representation functions $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ takes values in finite alphabet sets, and consequently induces quantizations of the original observation space \mathcal{X} . More precisely, every representation function $\mathbb{F}_i(\cdot)$ induces a measurable partition on \mathcal{X} , that we denote by $Q_{F_i} \subset \mathcal{F}_X$. In this scenario, the concept of embedded representation is better characterized by properties of the induced family of measurable partitions $\{Q_{F_i} : i = 1, \dots, n\}$ of \mathcal{X} . The following definition formalizes this point.

Definition 4: Let us consider our original space $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_X \times \mathcal{F}_Y))$ and a family of measurable representations $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$, taking values in finite alphabet sets $\{\mathcal{A}_i : i = 1, \dots, n\}$, i.e., $\mathbb{F}_i(\cdot) : (\mathcal{X}, \mathcal{F}_X) \rightarrow (\mathcal{A}_i, 2^{\mathcal{A}_i})$, with $|\mathcal{A}_i| < \infty$. The family of representations $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ is embedded if: $|\mathcal{A}_i| < |\mathcal{A}_{i+1}|$, for all $i \in \{1, \dots, n-1\}$ and $\forall j, i \in \{1, \dots, n\}$, $j > i$, there exists a function $\pi_{j,i}(\cdot) : \mathcal{A}_j \rightarrow \mathcal{A}_i$ such that

$$\mathbb{F}_i(x) = \pi_{j,i}(\mathbb{F}_j(x)), \quad \forall x \in \mathcal{X}.$$

Remark 4: Every representation function $\mathbb{F}_i(\cdot)$ produces a quantization of the original observation space given by $Q_{F_i} \equiv \{\mathbb{F}^{-1}(\{a\}) : a \in \mathcal{A}_i\} \subset \mathcal{F}_X$, where the embedded condition implies that: $\forall i, j$, $1 \leq i < j \leq n$, Q_{F_j} is a refinement of Q_{F_i} ⁶, (notation, $Q_{F_i} \ll Q_{F_j}$), and then, $Q_{F_1} \ll Q_{F_2} \ll \dots \ll Q_{F_n}$.

Finally, the following theorem extends the tradeoff between the Bayes error bound and estimation error for the case of representation functions inducing a sequence of embedded quantizations of the observation space. For these

⁶ \bar{Q} is a refinement of Q if, $\forall A \in Q$, $\exists \bar{Q}_A \subset \bar{Q}$ such that $A = \bigcup_{B \in \bar{Q}_A} B$.

results we also make use of the assumption of consistence for the empirical distributions across the sequence of embedded representations, which extends naturally from the continuous case presented in *Definition 3*.

THEOREM 3: Let $(X(u), Y(u))$ be the joint-observation random vector with joint probability $P_{X,Y}$. Let $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ be a family of embedded representation functions taking values in finite alphabet sets $\{\mathcal{A}_i : i = 1, \dots, n\}$. Considering the quantized observation family $\{X_i(u) \equiv \mathbb{F}_i(X(u)) : i = 1, \dots, n\}$ as observation rvs. for the Bayes classification rule, then the Bayes error bound satisfies:

$$L_{\mathcal{A}_{i+1}} \leq L_{\mathcal{A}_i} \quad (9)$$

$\forall i \in \{1, \dots, n-1\}$.

If in addition we have empirical probabilities $\hat{P}_{X_i,Y}$ on the family of representation spaces $\mathcal{A}_i \times \mathcal{Y}^7$ with conditional class probabilities, $\{\hat{P}_{X_i|Y}(\cdot|y) : i = 1, \dots, n\}$, consistent with respect to $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$, $\forall y \in \mathcal{Y}$, then the estimation error satisfies the following inequality:

$$\Delta g_{MAP}(\hat{P}_{X_{i+1},Y}) \geq \Delta g_{MAP}(\hat{P}_{X_i,Y}) \quad (10)$$

$\forall i \in \{1, \dots, n-1\}$. The proof is presented in *Appendix III*.

This result again shows the tradeoff between Bayes bound and estimation errors but for a family of embedded representations taking finite values. The tradeoff is obtained as a function of the cardinality of those spaces, and hence cardinality is the natural complexity indicator in this context.

The following proposition states the validity of the consistence condition for the important scenario when the empirical distribution is obtained using the Maximum Likelihood (ML) criterion. This result shows that THEOREM 3 is valid when the ML criterion — frequency counts — is used to characterize the empirical distributions in a family of finite alphabet embedded representation spaces.

Proposition 1: For a given amount of training data, iid realizations of $(X(u), Y(u))$, the ML estimator of $P_{X_i|Y}(\cdot|y)$, $\forall y \in \mathcal{Y}$ obtained in the range of a family of finite alphabet embedded representations $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$, per *Definition 4*, satisfies the consistence condition stated in *Definition 3*. The proof is presented in *Appendix IV*.

Before changing the topic of the discussion, it is important to emphasize that having a family of embedded representations, continuous or finite alphabet version, is not enough to show the results about the evolution of the estimation error across this embedded sequence of increasing complexity, presented in THEOREM 2 and 3. The additional necessary assumption is to have a consistent family of empirical distributions (see proofs of the theorems for details). This last condition is clearly a function of the learning methodology used for estimating the conditional class-observation distributions on the representation spaces.

Based on the results presented in this section, the performance of a Bayes classification approach, Eq.(4), is affected by two independent factors the intrinsic discrimination power of the observation space, quantified by the Bayes error bound, Eq.(2), and the estimation error, which provides a bound for the performance deviation when

⁷In this case we consider the power set of $\mathcal{A}_i \times \mathcal{Y}$ as the sigma field, and consequently we omit it.

the joint observation-class probability is estimated based on finite amount of training data, THEOREM 1. In close connection with this observation, THEOREM 2 and 3 formalize the tradeoff between the Bayes error bound and the estimation error in the process of increasing decreasing the complexity of the observation-feature space constrained to an embedded representation structure. As explained by Vasconcelos [1], this last result formally justifies the fact that in the process of doing dimensionality-cardinality reduction, better estimation of the underlying observation-class distribution is obtained, in the KLD sense, as a consequence of increasing the underlying Bayes error bound. Then, by constraining to a sequence of embedded representations there is one that minimizes the probability of error, the one that presents the optimal tradeoff between the Bayes error and the estimation error. This formally explains the “pickering phenomenon” observed in the operational error probability by exploring performances in a sequence of embedded spaces.

Consequently, the critical question is to find the optimal sequence of embedded representations along a large collection of observation spaces that will allow us to address the problem of optimal signal representation for classification. Recall that the final objective is to find the observation space that minimizes the probability of error of the empirical Bayes rule Eq.(1), considering the operational constraint that the true joint observation-class distribution is unavailable and is to be estimated with finite training data.

For doing that, we propose the characterization of a rate-distortion type of optimality criterion, where the formulation of a representation quality indicator or “fidelity criterion” and a cost or complexity-penalization function for every representation space is proposed, motivated by the tradeoff between Bayes bound and estimation error presented in this section. The general idea is to find the sequence of representation spaces which are solutions to the following optimization problem: for a given cost or penalization value, find the space whose representation quality for the decision problem is the best among all the ones that satisfy the above mentioned complexity-penalization constraint.

This formulation presents a natural analogy with the classical problem of characterizing the operational rate-distortion function associated with the region of rate-distortion achievable points, in lossy compression with a fidelity criterion [19], [25], [26], [32], [42]. In this case for a given rate constraint, which is the operational penalization for transmitting the quantized source through a channel, the problem is to find the “*quantization-representation*” scheme that minimizes a given distortion function, usually the MSE between the quantized source and the original one. In this scenario, there is a natural tradeoff between the representation quality of the quantization scheme and the rate needed for transmitting the quantized signal. A similar tradeoff is presented in the problem of signal representation for classification, where the representation quality is proportional to the complexity of the observation space, but also the cost in terms of number of training examples that we need to incur for having a good estimation of the joint observation-class distribution, per THEOREM 2 and 3.

IV. RATE-DISTORTION OPTIMALITY CRITERION

Let us consider the joint observation class label random phenomenon $(X(u), Y(u))$ with joint distribution $P_{X,Y}$ defined on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_{\mathcal{X}} \times \mathcal{F}_{\mathcal{Y}}))$. In this section, we will not put any assumption about the observation space \mathcal{X} and

we assume that $P_{X,Y}$ is given. In addition, let us consider a family of representation functions, denoted by \mathbb{D} , such that $\forall \mathbf{f}(\cdot) \in \mathbb{D}$, $\mathbf{f}(\cdot)$ is measurable from the original observation space $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ to a transform measurable space denoted by $(\mathcal{X}_f, \mathcal{F}_{\mathcal{X}_f})$, where $\mathcal{X}_f \equiv \{\mathbf{f}(x) : x \in \mathcal{X}\}$ is a function of the representation function $\mathbf{f}(\cdot)$. Applying $\mathbf{f}(\cdot)$ to $X(u)$, denoted by $X_f(u) \equiv \mathbf{f}(X(u))$, induces a lossy representation of the random observation $X(u)$, because $\mathbf{f}(\cdot)$ is not necessarily injective P_X almost surely. In fact, this will be our case of interest, because we are looking for simplification of the original observation phenomenon by using an optimality criterion.

Considering our application domain, the fidelity criterion for a given representation function $\mathbf{f}(\cdot) \in \mathbb{D}$ should be an indicator of the discriminative power of its induced observation space \mathcal{X}_f . Consequently, the probability of error of the Bayes classification rule, MAP decision rule Eq.(2), turns out to be the best candidate. This was defined as the Bayes error bound for \mathcal{X}_f , Eqn(2), and in this case is given by

$$L_{\mathcal{X}_f} = 1 - \mathbb{E}_{X_f} \left[\max_{i \in \mathcal{Y}} P_{Y|X_f}(i|X_f) \right]. \quad (11)$$

However $L_{\mathcal{X}_f}$ is very difficult to manipulate, in particular if we want to formalize an optimality criterion based on it⁸. We propose to use the mutual information [19], [41] which is strongly related with $L_{\mathcal{X}_f}$ and has the potential to provide a simpler description with respect to the induced observation-class distribution $P_{X_f,Y}$, which is the fundamental abstract object that characterizes the discriminative power of the observation space \mathcal{X}_f . In fact, an emblematic example of this general direction in the context of communication theory is the *second Shannon coding theorem* [19], [41], [44]. This well-known result formally proves a strong connection between the probability of error of an optimal decision theoretic approach for detecting communication symbols at certain rate, and an information theoretic quantity, the mutual information [17], [19], [26], which is function of the joint distribution between the source (class label) and the observation process at the receiver.

Mutual information quantifies the level of statistical dependency between two random variables, in this case $X(u)$ and $Y(u)$. More precisely, $I(X, Y)$ is the Kullback-Leibler divergence [18] between the joint distribution $P_{X,Y}$ and a distribution induced by the product of the marginals, $P_X \cdot P_Y$ [41]. Consequently, mutual information is a statistical indicator of how dissimilar is the joint distribution $P_{X,Y}$ for the case when X and Y are considered to be independent, the scenario where $X(u)$ is irrelevant in inferring $Y(u)$. Mutual information presents a strong relationship with the probability of error in statistical decision theory [19], [26] and consequently it turns out to be the natural candidate as an objective fidelity indicator for classification. The main reason in this application domain is because of Fano's inequality [19](*Chapter 2.11*), which characterizes a lower bound for the probability of error of any decision framework $g(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$ that tries to infer $Y(u)$ as a function of $X(u)$, by:

$$\begin{aligned} H(Y|X) &= H(E|X) + P_{X,Y}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : g(x) \neq y\}) \cdot H(Y|E = 1, X) \\ &\leq 1 + P_{X,Y}(g(X(u)) \neq Y(u)) \cdot H(Y) \end{aligned}$$

⁸There are some cases where the Bayes error bound can be approximated based on empirical data and used effectively as an optimality criterion. For instance, the empirical risk is used as fidelity indicator in regression and classification trees [2], [43].

where $E(u) \equiv \mathbb{1}_{\{Y(u) \neq g(X(u))\}}(u)$ is the binary error random variable. Knowing that mutual information can be written as the difference between unconditional and conditional class entropies, $I(Y, X) = H(Y) - H(Y|X)$ [19], we have that

$$P_e(g) \equiv P_{X,Y}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : g(x) \neq y\}) \geq \left(1 - \frac{I(X, Y) + 1}{H(Y)}\right), \quad (12)$$

where $I(X, Y)$ and $H(Y)$ are the mutual information between X and Y and the entropy of Y , respectively. This bound is true for any decision function $g(\cdot)$ and in particular for the optimal Bayes decision rule, Eq.(1), which is the tightest condition for this lower bound. Consequently, from Eq.(12), we have that $\forall \mathbf{f}(\cdot) \in \mathbb{D}$,

$$L_{\mathcal{X}_f} \geq \left(1 - \frac{I(X_f, Y) + 1}{H(Y)}\right). \quad (13)$$

Although this result just provides a lower bound for the probability of error, without ensuring any tightness condition, mutual information has been effectively used as an objective indicator in dimensionality reduction and feature extraction problems [21], [45]–[47] which strongly motivates its use in this context. Finally, for a given representation function $\mathbf{f}(\cdot) \in \mathbb{D}$, the fidelity criterion is given by:

$$I(\mathbf{f}) \equiv I(X_f, Y). \quad (14)$$

Regarding the penalization or cost function, this indicator depends on the dictionary of possible representations \mathbb{D} considered in the problem. In this respect, we will consider two important scenarios. If we restrict it to a family taking values in finite dimensional spaces, i.e., $\mathbf{f}(\cdot) \in \mathbb{D}$ then $\mathbf{f} : \mathcal{X} \mapsto \mathcal{X}_f = \mathbb{R}^k$ for some $k \in \mathbb{N}$ function of $\mathbf{f}(\cdot)$, the penalization can be associated with the dimensionality of \mathcal{X}_f . The justification comes from THEOREM 2, where it is shown that the estimation error increases by increasing the dimensionality of the representation space, in a sequence of dimensional embedded spaces. On the other hand, if we restrict \mathbb{D} to a family that quantizes the space \mathcal{X} into a finite number of bins, in other words the family of functions taking values in finite alphabets, i.e., $\mathbf{f}(\cdot) \in \mathbb{D}$, $\mathbf{f} : \mathcal{X} \mapsto \mathcal{A}_f = \{1, \dots, k\}$ for some $k \in \mathbb{N}$ function of $\mathbf{f}(\cdot)$, then the cost is naturally associated with the cardinality of \mathcal{A}_f . The justification for this consideration is again because of the evolution of the estimation error across a sequence of embedded quantization of the original observation space \mathcal{X} , THEOREM 3.

In both cases, the cost function is proportional to the amount of training data needed for having precise estimation of the joint induced distribution $P_{\mathbf{f}(X), Y}$, and consequently a natural operational penalization for the problem. For the rest of this section, let us generally denote $R(\mathbf{f})$ and $I(\mathbf{f})$ as the cost and fidelity indicators, respectively, $\forall \mathbf{f} \in \mathbb{D}$.

Definition 5: Let us consider the operational cost-fidelity region of achievable points on \mathbb{D} as

$$\mathcal{R}_{\mathbb{D}} \equiv \{(R(\mathbf{f}), I(\mathbf{f})) : \mathbf{f} \in \mathbb{D}\}. \quad (15)$$

Then we can naturally consider the operational fidelity-cost function on \mathbb{D} by

$$\hat{I}(D) \equiv \max_{\substack{\mathbf{f} \in \mathbb{D} \\ R(\mathbf{f}) \leq D}} I(\mathbf{f}). \quad (16)$$

$\hat{I}(D)$ characterizes the optimal boundary, from a signal representation point of view, of the achievable region $\mathcal{R}_{\mathbb{D}}$. In other words, for a given cost constraint $\hat{I}(D)$ is the best fidelity value achievable in \mathbb{D} .

The next subsection formalizes the problem of minimum probability of error signal representation (MPE-SR) for classification, and presents the role of the rate-distortion formulation for addressing it.

V. THE MINIMUM PROBABILITY OF ERROR SIGNAL REPRESENTATION (MPE-SR) PROBLEM

We begin by stating the minimum probability error (MPE) problem for a family of signal representations, under the Bayes classification approach. Let us consider $\{(x_i, y_i) : i = 1, \dots, N\}$ iid realizations of the joint observation-class phenomenon $(X(u), Y(u))$ with true distribution law $P_{X,Y}$ defined on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_X \times \mathcal{F}_Y))$. In addition, let us consider a family of measurable representation functions \mathbb{D} , where any $\mathbf{f}(\cdot) \in \mathbb{D}$ is defined in \mathcal{X} and takes values in \mathcal{X}_f . Every representation function $\mathbf{f}(\cdot)$ induces an empirical distribution $\hat{P}_{X_f, Y}$ on $(\mathcal{X}_f \times \mathcal{Y}, \sigma(\mathcal{F}_f \times \mathcal{F}_Y))$, based on the training data and implicit learning approach, and consequently the empirical Bayes classification rule is given by

$$\hat{g}_f(x) = \arg \max_{y \in \mathcal{Y}} \hat{P}_{X_f, Y}(x, y), \quad \forall x \in \mathcal{X}_f. \quad (17)$$

On the other hand, given that the family of transformations on \mathbb{D} are measurable, we have the original distribution $P_{X_f, Y}$ induced on the representation space $(\mathcal{X}_f \times \mathcal{Y}, \sigma(\mathcal{F}_f \times \mathcal{F}_Y))$, $\forall \mathbf{f}(\cdot) \in \mathbb{D}$. Consequently, restricted to the training data and the empirical Bayes classification rules, the MPE-SR problem reduces to:

$$\begin{aligned} \mathbf{f}^* &= \arg \min_{\mathbf{f} \in \mathbb{D}} \mathbb{E}_{X_f, Y} \left(\mathbb{1}_{\{(x, y) \in \mathcal{X}_f \times \mathcal{Y} : \hat{g}_f(x) \neq y\}}(X_f, Y) \right) \\ &= \arg \min_{\mathbf{f} \in \mathbb{D}} \mathbb{E}_{X, Y} \left(\mathbb{1}_{\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \hat{g}_f(\mathbf{f}(x)) \neq y\}}(X, Y) \right) \end{aligned} \quad (18)$$

where the last expected value is taken with respect to the underlying true joint distribution $P_{X,Y}$ on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_X \times \mathcal{F}_Y))$. Note that $\forall \mathbf{f}(\cdot) \in \mathbb{D}$, $\mathbb{E}_{X, Y} \left(\mathbb{1}_{\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \hat{g}_f(\mathbf{f}(x)) \neq y\}}(X, Y) \right) \leq L_{\mathcal{X}_f} \leq L_{\mathcal{X}}$, then the minimum probability of error tries to find the representation framework whose performance is the closest to $L_{\mathcal{X}}$, the fundamental bound for the problem.

Using the results presented in Section III-A, where an upper bound for the probability of error, $\mathbb{E}_{X_f, Y} \left(\mathbb{1}_{\{(x, y) \in \mathcal{X}_f \times \mathcal{Y} : \hat{g}_f(x) \neq y\}}(X_f, Y) \right)$, was derived as the sum of the risk of the empirical Bayes decision rule, Eq.(17), and an information theoretic indicator quantifying the deviation with respect to the Bayes error bound,

$$\mathbb{E}_{X_f, Y} \left(\mathbb{1}_{\{(x, y) \in \mathcal{X}_f \times \mathcal{Y} : \hat{g}_f(x) \neq y\}}(X_f, Y) \right) \leq \Delta g_{MAP}(\hat{P}_{X_f, Y}) + L_{\mathcal{X}_f}, \quad \forall \mathbf{f} \in \mathbb{D},$$

we can then follow the direction proposed in the empirical risk minimization (ERM) principle [16], where also results to control the deviation with respect to Bayes error bound were derived, to reduce Eq.(18) to the following optimization problem,

$$\mathbf{f}^* \approx \arg \min_{\mathbf{f} \in \mathbb{D}} \Delta g_{MAP}(\hat{P}_{X_f, Y}) + [L_{\mathcal{X}_f} - L_{\mathcal{X}}], \quad (19)$$

where we introduce the normalization factor $L_{\mathcal{X}}$ to make explicit that this optimization problem implies finding the optimal tradeoff between *approximation quality*, $L_{\mathcal{X}_f} - L_{\mathcal{X}}$, and *estimation error*, $\Delta g_{MAP}(\hat{P}_{X_f, Y})$.

In addition in Section III, we show that the performance of the empirical Bayes classification rule $\hat{g}_f(\cdot)$ is influenced by the “*Bayes bound - estimation*” tradeoff in the process of considering an embedded representation

sequence of increasing complexity. Then, as presented in Eq(19) the MPE-SR can be formulated as a complexity regularized optimization problem whose objective function consists of a weighted combination of a fidelity criterion, reflecting the Bayes error bound, and a cost term, penalizing the complexity of the representation scheme, reflecting the estimation error. For practical implementation, our fidelity and penalization functions presented in Section IV come into play and the problem can be generally approximated by the following complexity regularized fidelity criterion:

$$\mathbf{f}^*(\lambda) = \arg \min_{\mathbf{f} \in \mathbb{D}} \Psi(I(\mathbf{f})) + \lambda \cdot \Phi(R(\mathbf{f})), \quad (20)$$

where considering the tendency of the fidelity-cost indicators, $\Psi(\cdot)$ should be a strictly decreasing real function and $\Phi(\cdot)$, a strictly increasing function from \mathbb{N} to \mathbb{R} . Noting that the real dependency between Bayes and estimation error in terms of our new fidelity complexity values, $I(\mathbf{f})$ and $R(\mathbf{f})$, is hidden and, furthermore, problem dependent. In this scenario, Ψ , Φ and λ provide degrees of freedom for approximating the solution of the MPE-SR problem, Eq.(19). It is interesting to note that independent of those degrees of freedom, the optimal solution $\mathbf{f}^*(\lambda)$ resides in the sequence of representations which are solutions to the rate-distortion problem presented in Section IV, Eq.(16). More precisely,

$$\mathbf{f}^*(\lambda) = \arg \min_{\mathbf{f} \in \{\mathbf{f}_k^* : k \in K(\mathbb{D})\}} \Psi(I(\mathbf{f})) + \lambda \cdot \Phi(R(\mathbf{f})), \quad (21)$$

where $\{\mathbf{f}_k^* : k \in K(\mathbb{D})\} \subset \mathbb{D}$ is the family of representation functions that are solutions of the rate-distortion problem, i.e.,

$$\mathbf{f}_k^* = \arg \max_{\substack{\mathbf{f} \in \mathbb{D} \\ R(\mathbf{f}) \leq k}} I(\mathbf{f}), \quad (22)$$

$\forall k \in K(\mathbb{D})$, where $K(\mathbb{D}) \equiv \{R(\mathbf{f}) : \mathbf{f} \in \mathbb{D}\} \subset \mathbb{N}$.

Note that the cardinality of $K(\mathbb{D})$ could be significantly smaller than $|\mathbb{D}|$ and as a result, the domain of solutions of the original problem can be significantly reduced. Then, the MPE-SR solution can be restricted to the solution of the operational rate-distortion family $\{\mathbf{f}_k^* : k \in K(\mathbb{D})\}$. Finally, the minimum empirical risk criterion, for instance using cross validation, can be the final decision step for solving Eq.(21), as it has been used successfully for addressing a similar problem in the context of regression and classification trees [2], [43]. Finally, this formulation requires us to compute $I(\mathbf{f})$ based on the true joint class-observation distributions, however for practical purposes this can be approximated based on its empirical counterpart $\hat{I}(\mathbf{f})$, using $\hat{P}_{X_f, Y}$ in Eq.(14).

Remark 5: We could in general consider the MPE-SR under the unconstrained family of measurable representations, \mathbb{D} , where finding the solution for the rate-distortion (R-D) problem, Eq.(22), would again be intractable for any practical purpose. However, as in lossy source coding, this scenario is conceptually interesting because it characterizes the theoretical rate-distortion boundary for the problem [19], [26]. The extension of an equivalent *condign theorem* that would allow an alternative characterization of this theoretical bound is an open problem in this scenario. We think that it is not possible to extend it from classical results in rate-distortion theory for lossy compression [19], [25], [32] because we are dealing with a completely different cost function.

The MPE-SR formulation presented in this section was motivated by the problem of binary classification trees introduced by Breiman, Friedman, Olshen and Stone [2], where a similar formulation for finding the optimal MPE binary tree-indexed quantization was presented, [43]. Moreover, in the next section we will revisit this classic signal representation problem, and show that it can be considered a particular case of the formulation presented in this paper. We will show that the well known CART binary tree pruning algorithms [2], [33], [43], address a similar complexity regularized optimization problem where the rate-distortion formulation is implicitly used.

VI. REVISITING RESULTS FOR THE OPTIMAL BINARY CLASSIFICATION TREES: CART PRUNING ALGORITHMS

Putting some restriction on the raw observation \mathcal{X} and more importantly on the family of representation functions \mathbb{D} , we can characterize different sub-problems associated with the general rate-distortion formulation presented in Section V. For instance, considering $\mathcal{X} = \mathbb{R}^K$, a finite dimension Euclidean space, and the family of representation functions taking values in finite alphabet spaces — where the cost is proportional to the cardinality of the induced space — the problem reduces to finding the family of optimal vector quantizations. This problem can be considered as a variation of the problem of lossy compression with a source degraded by noise [48]–[51], with the main differences in our problem being that the fidelity criterion is the probability of error instead of the mean square error, the standard fidelity criterion used in lossy compression, and the cost function is the cardinality of the quantization instead of the rate of the quantized source.

If we further restrict the problem to the family of vector quantization with binary tree indexed structure [49], [52], [53], then the problem reduces to finding an optimal binary classification tree topology, where pruning tree algorithms originally proposed by Breiman, Friedman, Olshen and Stone (BFOS) [2]⁹, can be shown to address an instance of the complexity regularized problem presented in Section V. In this section we formally present the details of this connection, and in particular how the Bayes error bound estimation error tradeoff is valid in this context. This justifies the formulation of a complexity regularized optimization problem, Eqs. (20) and (21), and the fidelity-cost formalization for the MPE-SR problem.

Let us introduce some basic terminology to formalize the problem in this scenario. The interested reader is referred to [2], [54] for a more systematic exposition of the CART algorithms. Using Breiman et al conventions [2], a tree T is represented by a collection of nodes, with implicit left and right mappings, reflecting the parent-child relationship among nodes. Hence, those mapping functions represent an acyclic connected graph. T is a *rooted binary tree* if every nonterminal node has two descendants. Let us denote $t_{root} \in T$ as the root of the tree, and $\mathcal{L}(T) \subset T$ the sub-collection of leaves or terminal nodes, nodes that do not have a descendent. We define the norm of the tree, $|T|$, as the cardinality of $\mathcal{L}(T)$, which will reflect the penalization or cost function later on. Let $S \subset T$, and if $t_{root} \in S$ and S is a rooted binary tree by itself, then we say that S is a pruned version of T , and we denote this relationship by $S \ll T$. For the rest of this section we will consider \mathbf{T}_{full} as the collection of all nodes associated with the maximal rooted binary tree and $m \equiv |\mathbf{T}_{full}|$.

⁹The seminal work of Breiman et al [2] addresses the more general case of classification and regression trees (CART), where for the context of this work we just highlight results concerning the classification part.

The tree structure is used to index a family of vector quantizations for \mathcal{X} . In order to formalize this idea, we can consider that every node $t \in \mathbf{T}_{full}$ has associated a measurable subset $\mathcal{X}_t \subset \mathcal{X}$, such that $\mathcal{X}_{t_{root}} = \mathcal{X}$ and if t_1 and t_2 are the direct descendants of a nonterminal node t , we then have that

$$\begin{aligned}\mathcal{X}_t &= \mathcal{X}_{t_1} \cup \mathcal{X}_{t_2}, \\ \mathcal{X}_{t_1} \cap \mathcal{X}_{t_2} &= \phi.\end{aligned}$$

Therefore, it is straightforward to show that any rooted binary tree $T \ll \mathbf{T}_{full}$ induces a measurable partition of the observation space given by $\mathcal{V}_T = \{\mathcal{X}_t : t \in \mathcal{L}(T)\}$, where more importantly, if $T_1 \ll T_2$ then \mathcal{V}_{T_2} is a refinement of \mathcal{V}_{T_1} . With this concept in mind, we can define a pair of tree indexed representations by $[T, \mathbf{f}_T(\cdot)] \forall T \ll \mathbf{T}_{full}$, with $\mathbf{f}_T(\cdot)$ being a measurable representation function from $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ to $\mathcal{L}(T)$, such that

$$\mathcal{X}_t = \mathbf{f}_T^{-1}(\{t\}), \quad \forall t \in \mathcal{L}(T). \quad (23)$$

Hence, $\mathbf{f}_T(\cdot)$ induces the previously defined measurable partition \mathcal{V}_T on $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$. Finally, the family of tree indexed representation is given by $\mathbb{D} = \{\mathbf{f}_T(\cdot) : T \ll \mathbf{T}_{full}\}$.

The next natural step is to define the Bayes classification rule, Eq. (1), associated with a given representation in \mathbb{D} . For that, let us consider again that we know the observation class probability measure $P_{X,Y}$ on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_{\mathcal{X}} \times \mathcal{F}_{\mathcal{Y}}))$. We follow a similar convention presented in [43], by defining a classification tree as a triple $[T, \mathbf{f}_T(\cdot), g_T(\cdot)]$, where $g_T(\cdot)$ from $\mathcal{L}(T)$ to \mathcal{Y} denotes the Bayes classification rule to infer $Y(u)$ based on the quantized observation random variable $X_{\mathbf{f}_T}(u) \equiv \mathbf{f}_T(X(u))$, which is given by:

$$g_T(t) = \arg \max_{y \in \mathcal{Y}} P_{X_{\mathbf{f}_T}, Y}(t, y),$$

$\forall t \in \mathcal{L}(T)$.

The probability of error or the average risk associated with the Bayes classification tree $[T, \mathbf{f}_T(\cdot), g_T(\cdot)]$ is:

$$\begin{aligned}R(T) &\equiv \mathbb{P}(\{u \in \Omega : g_T(X_{\mathbf{f}_T}(u)) \neq Y(u)\}) \\ &= P_{X_{\mathbf{f}_T}, Y}(\{(t, y) \in \mathcal{L}(T) \times \mathcal{Y} : g_T(t) \neq y\}).\end{aligned} \quad (24)$$

Breiman et al [2] (*Chapter 9*) show that this function can be written as an additive non-negative function of the terminal nodes of T , more precisely,

$$R(T) = \sum_{t \in \mathcal{L}(T)} R(t), \quad (25)$$

where for our target 0-1 decision rule $R(t)$ is given by:

$$R(t) = \mathbb{P}(X_{\mathbf{f}_T}(u) = t) \cdot \left(1 - \max_{y \in \mathcal{Y}} P_{Y|X_{\mathbf{f}_T}}(y|t)\right). \quad (26)$$

It is not difficult to show that if we have a sequence of embedded trees, $T_1 \ll T_2 \ll \dots \ll T_n$, they induce a measurable sequence of embedded representations $\{\mathbf{f}_{T_1}, \dots, \mathbf{f}_{T_n}\}$, in the sense presented in *Definition 4*. Moreover, it is direct from **THEOREM 3**, noting that the Bayes error bound $L_{\mathcal{L}(T)}$ is equivalent to the notion of average risk $R(T)$ presented in Eq.(24), that if $\bar{T} \ll T$ then $R(T) \leq R(\bar{T})$, which was proved in [2] (*Theorem 9.4*). In the

context of this exposition, this result comes naturally from the fact that \mathcal{V}_T is a refinement of $\mathcal{V}_{\bar{T}}$, per THEOREM 3.

Under the assumption that $P_{X,Y}$ is available, the best performance in \mathbb{D} is obtained for the finest representation, i.e., for the classification tree $[\mathbf{T}_{full}, \mathbf{f}_{\mathbf{T}_{full}}(\cdot), g_{\mathbf{T}_{full}}(\cdot)]$. However, the more interesting and realistic case is when we only have a finite amount of training data, $D_N = \{(x_i, y_i) : i = 1, \dots, N\}$ iid realizations of $(X(u), Y(u))$, and under this constraint, we want to address the MPE-SR problem formulated in Section V. In this case, the maximum likelihood (ML) empirical distribution $\hat{P}_{X_{\mathbf{T}_T}, Y}$ is considered, which reduces to a family of classification trees $[T, \mathbf{f}_T(\cdot), \hat{g}_T(\cdot)]$ where $\hat{g}_T(\cdot)$ is the empirical Bayes decision corresponding to the majority vote decision rule [2].

We will show that the “*Bayes bound - estimation error*” tradeoff holds for a sequence of embedded representations in \mathbb{D} , where we need to show that the empirical class conditional distributions are consistent across the given embedded sequence. For the rest of this exposition we will just consider the tree index T for referring to representation function $\mathbf{f}_T(\cdot)$ and the implicit empirical Bayes classification tree $[T, \mathbf{f}_T(\cdot), \hat{g}_T(\cdot)]$, depending on the context.

Proposition 2: Let us take a sequence of embedded trees $T_1 \ll T_2 \ll T_3, \dots, \ll T_k$, subsets of \mathbf{T}_{full} . Considering the Bayes classification rules for this family of representations, Eq.(24), we have that $\forall i \in \{1, \dots, n-1\}$,

$$R(T_{i+1}) \leq R(T_i). \quad (27)$$

In addition, for a given training data D_N , and the corresponding family of empirical classification trees, $[T_i, \mathbf{f}_{T_i}(\cdot), \hat{g}_{T_i}(\cdot)]$, $i \in \{1, \dots, k\}$, the estimation error of the empirical decision with respect to the Bayes classification rule THEOREM 1, denoted by $\Delta g(\hat{P}_{X_{\mathbf{T}_i}, Y})$, satisfies:

$$\Delta g(\hat{P}_{X_{\mathbf{T}_i}, Y}) \leq \Delta g(\hat{P}_{X_{\mathbf{T}_{i+1}}, Y}), \quad \forall i \in \{1, \dots, n-1\}. \quad (28)$$

Proof: We have developed all the machinery to prove this result. We know that the family of representations $\{\mathbf{f}_{T_1}(\cdot), \dots, \mathbf{f}_{T_n}(\cdot)\}$ is embedded, where by Proposition 1 the induced empirical distributions — conditional class probabilities — are consistent with respect to the embedded representation family. Consequently, the result extends directly from THEOREM 3. ■

As a result, the complexity regularized optimization criterion considered for addressing the MPE-SR problem, Section V, is justified. Note that in this case, we can take advantage of the tree embedded structure of the family of quantizers, in particular the additive property of probability of error, Eqs.(25) and (26), to use directly the probability of error (its empirical version for practical purpose) and cardinality as the fidelity and complexity indicators, Eq.(29), which in fact are additive tree functionals [49].

The practical solution to this problem is the well known CART pruning algorithm [2], which addresses and finds an algorithmic solution for the following optimization problem,

$$T_n^*(\alpha) = \arg \min_{T \ll \mathbf{T}_{full}} \hat{R}(T) + \alpha \cdot |T|, \quad (29)$$

where $\hat{R}(T)$, Eq.(30), is the empirical risk obtained from the training set, acting as the fidelity criterion in this context, and $\alpha \cdot |T|$ is the penalization term.

$$\hat{R}(T) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{(t,y) \in \mathcal{L}_T \times \mathcal{Y} : \hat{g}_T(t) \neq y\}}(\mathbf{f}_T(x_i), y_i) \quad (30)$$

Breiman et al [2] (*Chapter 10*) used the additivity of the fidelity, Eq.(25), and the cost function, $\lambda |T|$, to formulate a dynamic programming approach that solves the MPE pruned tree problem, $T_n^*(\alpha)$, in $O(|\mathbf{T}_{full}|)$. Moreover, they proved that there is a sequence of optimal embedded representations, denoted by $\mathbf{T}_{full} = T_1^* \gg T_2^* \gg \dots \gg T_m^* = \{t_{root}\}$, which are the solutions of Eq.(29) for all possible values of the complexity weight $\alpha \in \mathbb{R}^+$. More precisely,

$\exists \alpha_0 = 0 < \alpha_1 < \dots < \alpha_m = \infty$, and $\forall i \in \{1, \dots, m\}$, such that,

$$T_n^*(\alpha) = T_i^*, \quad \forall \alpha \in [\alpha_{i-1}, \alpha_i). \quad (31)$$

This is the result that connects the optimal tree pruning problem with the solutions for the rate-distortion problem, as Scott [33] had recently pointed out. The reason is that this family of optimal embedded sequences is the solution to the rate distortion problem, Eq.(21), which is given here by:

$$T_j^* = \arg \min_{\substack{T \ll \mathbf{T}_{full} \\ |T| \leq m-j-1}} \hat{R}(T) \quad (32)$$

$\forall j \in \{1, \dots, m\}$.

Scott coined the solution of Eq.(32) as the *minimum cost tree*, where a general algorithm to solve it in $O(|\mathbf{T}_{full}|^2)$ was presented [33]. Also the connection of this rate-distortion solution with a more general complexity regularized objective optimality criterion was presented, where the cost term $\alpha \cdot |T|$ is substituted for a general sized-based penalty $\alpha \cdot \Phi(|T|)$, where $\Phi(\cdot)$ is a non-decreasing function. Geometric algorithms based on the characterization of the operational rate-distortion boundary, Eq.(32), were presented for finding explicitly $\alpha_0 < \alpha_1 < \dots < \alpha_m$ as in Eq.(31) but for the general size-based penalty scenario. As far as the authors know, Scott's work is the first one that formally presents connections between the CART complexity regularized pruning problem with general sized-based penalty and the solution of a rate-distortion formulation, Eq.(32), perhaps the classical embedded solution presented in Eq.(31) implicitly addresses it for the case of additive penalties.

Once we have the family of optimal representation solutions $T_1^* \gg T_2^* \gg \dots \gg T_m^*$, the question reduces to finding the optimal α^* that minimizes $R(T^*(\alpha))$ [33], [43]. Cross-validation has been used to address this problem, where the average risk can be evaluated, restricted to this optimal representation family.

CART algorithms also characterize the family of representations based on the training data, which we consider given in our formulation, the dictionary \mathbb{D} . It uses a greedy algorithm that hierarchically partitions the observation space in an iterative fashion, such that in each step the maximum performance gain of the empirical Bayes rule is achieved [2], taking advantage of the additivity of $\hat{R}(\cdot)$. This process is conducted until the point at which the induced partition over-fits the data — the number of bins is close to the number of training examples — and

consequently, $R(\mathbf{T}_{full})$ is mainly dominated by the estimation error. This was the main justification for addressing the optimal tree pruning problem that we revisited in this section.

In sum, this work provides a well founded theoretical justification for the complexity regularized formulation of the optimal tree pruning problem, Eq.(29), based on the underlying tradeoff between the main two sources of degradation that affect the probability of error of the learning decisions made by the Bayes decision framework. Moreover, the indexed family of binary-tree representations, \mathbb{D} , have a rich embedded structure, in the sense presented in Section III, that provides an even stronger justification for the complexity regularized fidelity criterion, Eq.(29).

VII. REDUCING THE MPE-SR PROBLEM TO A LINEAR DISCRIMINANT ANALYSIS PROBLEM

Let us consider again the general formulation of the MPE-SR problem presented in *Section V*, with a finite dimensional observation space, $\mathcal{X} = \mathbb{R}^K$, and the family of linear transformations as the dictionary of feature representations $\mathbb{D} = \{\mathbf{f} : \mathbb{R}^K \rightarrow \mathbb{R}^m : \mathbf{f} \text{ linear}, m \leq K\}$. An element $\mathbf{f} \in \mathbb{D}$ can be univocally represented by a matrix $\mathbf{A} \in \mathbb{R}(m, K)^{10}$. In particular without loss of generality, we can restrict \mathbb{D} to the family of full rank matrices, i.e., the matrices with linear independent rows. For the rest of this section, we follow the modeling assumptions proposed by Padmanabhan et al [21] where the problem of dimensionality reduction is addressed under some parametric assumption for the conditional class observation distribution and mutual information is used as the objective indicator. If we consider that the conditional class probability follows a multivariate Gaussian distribution, then:

$$\begin{aligned} p_{X|Y}(\cdot|y) &= \mathcal{N}(\cdot, \mu_y, \Sigma_y) \\ p_X(\cdot) &= \sum_{y \in \mathcal{Y}} \mathbb{P}(Y(u) = y) \cdot \mathcal{N}(\cdot, \mu_y, \Sigma_y), \end{aligned} \quad (33)$$

where $\mathcal{N}(\cdot, \mu, \Sigma)$ is a multivariate Gaussian pdf with mean μ and covariance matrix Σ .

Under this assumption we have that any linear combination of $X(u)$ also has a Gaussian class conditional pdf and consequently we have that $\mathbf{A}X(u)$ satisfies:

$$\begin{aligned} p_{\mathbf{A}X|Y}(\cdot|y) &= \mathcal{N}(\cdot, \mathbf{A}\mu_y, \mathbf{A}\Sigma_y\mathbf{A}^\dagger) \\ p_{\mathbf{A}X}(\cdot) &= \sum_{y \in \mathcal{Y}} \mathbb{P}(Y(u) = y) \cdot \mathcal{N}(\cdot, \mathbf{A}\mu_y, \mathbf{A}\Sigma_y\mathbf{A}^\dagger). \end{aligned} \quad (34)$$

Considering a finite amount of training data $\{(x_i, y_i) : i = 1, \dots, N\}$ and maximum likelihood (ML) estimation techniques [14], the empirical distributions $\{\hat{p}_{X|Y}(\cdot|y) : y \in \mathcal{Y}\}$ and $\hat{p}_X(\cdot)$ are Gaussian mixtures and Gaussian, respectively, characterized by the empirical mean and covariance matrices given by:

$$\hat{\mu}_y = \frac{1}{N_y} \sum_{i=1}^N \mathbb{1}_{\{y\}}(y_i) \cdot x_i \quad (35)$$

$$\hat{\Sigma}_y = \frac{1}{N_y} \sum_{i=1}^N \mathbb{1}_{\{y\}}(y_i) \cdot (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\dagger, \quad (36)$$

¹⁰ $\mathbb{R}(m, n)$ represents the collection of $m \times n$ matrices with entries in \mathbb{R} .

with $N_y = |\{1 \leq i \leq N : y_i = y\}|$, $\forall y \in \mathcal{Y}$.

Proposition 3: Consider a family of full rank linear representations denoted by $\mathbf{A}_1, \dots, \mathbf{A}_n$, taking values in a sequence of spaces $\{\mathbb{R}^{k_1}, \dots, \mathbb{R}^{k_n}\}$ with $0 < k_1 < k_2 < \dots < k_n \leq K$. In addition, let us assume that the sequence of transformations is dimensionally embedded, *Definition 2*, i.e., $\forall j, i, j > i$ there exists $B_{j,i} \in \mathbb{R}(k_j, k_i)$, such that $\mathbf{A}_i = B_{j,i} \cdot \mathbf{A}_j$. Under the Gaussian parametric assumption for the class conditional distributions, then the empirical sequence of class conditional pdfs $\{\hat{p}_{\mathbf{A}_i X|Y}(\cdot|y) : i = 1, \dots, n\}$, estimated across $\{\mathbb{R}^{k_1}, \dots, \mathbb{R}^{k_n}\}$ by the ML estimation criterion, characterize a sequence of consistent probability measures with respect to $\{\mathbb{R}^{k_1}, \dots, \mathbb{R}^{k_n}\}$, in the sense presented in *Definition 3*. The proof is presented in *Appendix V*.

This result is very important because it formally extends THEOREM 2, the tradeoff between the Bayes error bound and estimation error, for the case of any sequence of full rank linear transformations $\mathbf{A}_1, \dots, \mathbf{A}_n$ from \mathbb{R}^K to a sequence of spaces with the deterministic embedded structure stated in Proposition 3. This result provides justification to address the MPE-SR problem, under the modeling assumption presented in this section and the dictionary of full rank linear transformations, using the rate-distortion approach presented in *Section V*. More precisely, the solution of the MPE-SR problem resides in the solution of the following rate-distortion problem:

$$\mathbf{A}_k^* = \arg \max_{\mathbf{A} \in \mathbb{R}(k, K)} I(\mathbf{A}) \quad (37)$$

$\forall k \in \{1, \dots, K\}$, where $I(\mathbf{A})$ is the mutual information between rvs. $\mathbf{A}X(u)$ and $Y(u)$.

For addressing this optimization problem we follow the direction proposed in [21] for simplifying the maximum mutual information problem, Eq.(37). We have that,

$$\begin{aligned} I(\mathbf{A}) &= I(\mathbf{A}X(u), Y(u)) \\ &= H(\mathbf{A}X(u)) - H(\mathbf{A}X(u)|Y(u)) \end{aligned} \quad (38)$$

where under the Gaussian assumption and considering $\mathbf{A} \in \mathbb{R}(k, K)$, it follows that [19],

$$H(\mathbf{A}X(u)|Y(u) = y) = \frac{k}{2} \log(2\pi) + \frac{1}{2} \log(|\mathbf{A}\Sigma_y \mathbf{A}^\dagger|) + \frac{1}{2}.$$

Given that $\mathbf{A}X(u)$ has a Gaussian mixture distribution, a closed form expression is not available for the differential entropy. Padmanabhan et al [21] proposed to use an upper bound based on the well known fact that the Gaussian law maximizes the entropy under second moment constraints [19]. Then, denoting $\Sigma \equiv \mathbb{E}(X(u)X(u)^\dagger) - \mathbb{E}(X(u))\mathbb{E}(X(u))^\dagger$, we have that

$$H(\mathbf{A}X(u)) \leq \frac{k}{2} \log(2\pi) + \frac{1}{2} \log(|\mathbf{A}\Sigma \mathbf{A}^\dagger|) + \frac{1}{2}$$

and then

$$\begin{aligned} I(\mathbf{A}) &\leq \frac{1}{2} \log(|\mathbf{A}\Sigma \mathbf{A}^\dagger|) - \frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{P}(Y(u) = y) \cdot \log(|\mathbf{A}\Sigma_y \mathbf{A}^\dagger|) \\ &= \frac{1}{2} \log \left[\frac{|\mathbf{A}\Sigma \mathbf{A}^\dagger|}{\prod_{y \in \mathcal{Y}} |\mathbf{A}\Sigma_y \mathbf{A}^\dagger|^{\mathbb{P}(Y(u)=y)}} \right]. \end{aligned} \quad (39)$$

Then the rate distortion problem reduces to

$$\mathbf{A}_k^* = \arg \max_{\mathbf{A} \in \mathbb{R}^{(k, K)}} \log \left[\frac{|\mathbf{A} \Sigma \mathbf{A}^\dagger|}{\prod_{y \in \mathcal{Y}} |\mathbf{A} \Sigma_y \mathbf{A}^\dagger|^{\mathbb{P}(Y(u)=y)}} \right]. \quad (40)$$

In practice, we need to address Eq.(40) based on empirical distributions estimated with a finite amount of training data. Then, it reduces to characterizing the empirical class conditional covariance matrices $\hat{\Sigma}_y$, Eq(36), and the unconditional empirical matrix $\hat{\Sigma}$ that can be written as $\hat{\Sigma}_w + \hat{\Sigma}_b$ [21],

$$\begin{aligned} \hat{\Sigma}_w &\equiv \sum_{y \in \mathcal{Y}} \hat{P}_Y(\{y\}) \cdot \hat{\Sigma}_y \\ \hat{\Sigma}_b &\equiv \sum_{y \in \mathcal{Y}} \hat{P}_Y(\{y\}) \cdot (\hat{\mu} - \hat{\mu}_y)(\hat{\mu} - \hat{\mu}_y)^\dagger \end{aligned} \quad (41)$$

where $\hat{\Sigma}_w$ and $\hat{\Sigma}_b$ are the well known between-class and within-class scatter matrices used in linear discriminant analysis [14] and $\hat{\mu}$ is the unconditional empirical mean. As pointed out in [21], under the additional assumption that class conditional covariance matrixes are equivalent, the problem reduces to

$$\mathbf{A}_k^* = \arg \max_{\mathbf{A} \in \mathbb{R}^{(k, K)}} \log \left[\frac{|\mathbf{A} \hat{\Sigma} \mathbf{A}^\dagger|}{|\mathbf{A} \hat{\Sigma}_y \mathbf{A}^\dagger|} \right],$$

which is exactly the objective function used for finding the optimal linear transformation used in multiple discriminant analysis (MDA), the case $k = 1$ being the Fisher linear discriminant analysis problem [14]. Then under the Gaussian parametric assumption about the class conditional distribution, the MDA problem approximates the solution of the rate-distortion problem by optimizing and upper bound of the mutual information.

VIII. FINAL DISCUSSION AND CONNECTIONS WITH EMPIRICAL RISK MINIMIZATION (ERM) AND STRUCTURAL RISK MINIMIZATION (SRM) INDUCTIVE PRINCIPLE

The MPE-SR formulation presents some interesting conceptual connection with the theory of *empirical risk minimization* (ERM) for pattern recognition. Comprehensive exposition of this learning principle can be found in [15], [16], [35]; a good survey emphasizing recent results in pattern classification can be found in [20]. The ERM principle considers a class \mathbb{C} of classifiers — a subset of the set of measurable functions from \mathcal{X} to \mathcal{Y} — and naturally formalizes the learning problem as finding the decision rule $g(\cdot)$ in \mathbb{C} that minimizes the empirical risk $\hat{R}(g)$, Eq.(42), based on a finite amount of training data, $\{(x_i, y_i) : i = 1, \dots, N\}$: iid realization of the observation and class random phenomenon $(X(u), Y(u))$ taking values in $\mathcal{X} \times \mathcal{Y}$.

$$\hat{R}(g) \equiv \frac{1}{N} \cdot \sum_{i=1}^N \mathbb{1}_{\{g(x_i) \neq y_i\}} \quad (42)$$

In this formulation the observation feature space \mathcal{X} is fixed and the learning problem reduces to finding the decision rule in \mathbb{C} that minimizes the empirical risk, $g_N^*(\cdot) \equiv \arg \min_{g \in \mathbb{C}} \hat{R}(g)$. Formal results have been derived to show consistency of the learning principle in the classical asymptotic statistical sense as the number of samples goes to infinity. Furthermore, uniform bounds in \mathbb{C} for the rate of convergence of empirical risk $\hat{R}(g)$ to the expected

risk $R(g) \equiv \mathbb{P}(\{g(X(u)) \neq Y(u)\})$ have been derived as a function of a combinatorial notion of the complexity for the family of decision functions \mathbb{C} , the VC dimension [55], [56].

The notion of VC dimension is particularly crucial in the development of this theory because it allows controlling the generalization ability of the learning principle (how far $R(\hat{g}_N^*)$ can be from the actual minimal risk decision in \mathbb{C} , $\inf_{g \in \mathbb{C}} R(g)$, independent of the joint distribution of $(X(u), Y(u))$). This is particularly crucial when dealing with sample sizes which are small relative to the VC dimension of the class of functions \mathbb{C} , and consequently the gap between empirical and average risk turns out to be significant. This last scenario presents the formal justification to address the learning problem as a complexity regularized optimization problem, where in one hand we have a fidelity function (empirical risk) and on the other some notion of complexity, in which this learning theory is explicitly a function of the VC dimension for a given family of classifiers and the number of training examples [15], [16]. The *structural risk minimization* (SRM) principle was aimed at formalizing this regularization problem which was proposed to solve the optimal tradeoff between *fidelity* and *complexity* for a given amount of training data in a sequence of classifier families, $\mathbb{C}_1 \subset \mathbb{C}_2 \subset \dots$, with a *structure* of increasing VC dimensions.

At this point it is interesting to discuss some natural analogies and differences with the formulation of the MPE-SR problem. First, the MPE-SR makes use of the Bayes decision approach as a way to define the optimal decision rule based on estimated empirical distributions, and the domain to address the MPE learning problem is with respect to a family of feature representations. Consequently the learning principle is conceptually different than the one used in the SRM that uses empirical risk minimization and also the way to decide the optimal decision rule having the degree of freedom in a family of classifiers. Then, results from the ERM inductive principle cannot be directly extended into the Bayes decision learning approach.

Second, as in the ERM learning approach the MPE-SR approach provides an upper bound for measuring the generalization abilities of the empirical Bayes rule with respect to the optimal Bayes rule. More precisely, the deviation of the risk of the empirical Bayes rule with respect to the Bayes error bound — estimation error— is a function of an information theoretic quantity, the average KLD between the involved class conditional distributions, Section III-A. In this respect, it is not possible to characterize a universal closed-form expression as the one obtained in ERM theory. However, under some parametric assumption like that of multivariate Gaussian distribution, generalized Gaussian distribution, or Gaussian mixture models, KLD closed form expressions or KLD upper bound closed form expressions are available [19], [57], which would allow us to find distribution dependent bounds to analyze the rate of convergence of performance of the empirical Bayes rule to the Bayes error bound, as a function of the number of training points and the dimensionality of the feature space. This issue is interesting to address, in particular considering the fact that this family of parametric models has been used extensively under the Bayes decision approach [14]. On the other hand, in the MPE-SR the notion of complexity is directly associated with common engineering indicators, cardinality and dimensionality of the feature space depending on the family of representations considered in the problem. This is not the case for the ERM principle where the VC dimension is an abstract concept and potentially difficult to characterize for a given family of classifiers.

Third, results that formally present the tradeoff between Bayes error and estimation error across sequences of

embedded representation presented in Section III-B, and the main motivation to address the MPE-SR problem as a complexity regularized optimization problem has an equivalent counterpart in the SRM inductive principle. This explains why the two frameworks address similar tradeoffs between complexity and fidelity for finding the minimum probability of error decision rule constrained to a finite amount of data.

Regarding important issues of practical implementation, the MPE-SR needs to address the complexity regularized optimization problem or equivalently find a solution to the rate-distortion problem. Given the empirical probability of error as fidelity, in most of the cases this indicator does not have any closed-form expression. Empirical mutual information turns to be a naturally attractive candidate in particular considering some family of parametric models and potential embedded structure of feature representations, which is the motivation of the formulation presented in Section IV. In this direction this paper presents two emblematic learning scenarios that show how this principle can be practically implemented: one under some parametric assumptions and approximations for the case of a finite dimensional feature family (Section VII), and the other considering a family of vector quantizations with a strong tree embedded structure, where the induced combinatorial problem can be solved using dynamic programming (DP) techniques, (Section VI). The ERM inductive principle has the same issues for addressing the empirical risk minimization problem. In this case some approximations based on discriminant cost functions have been considered which allow to address the optimization problem for particular families of classifiers. Examples of those practical learning frameworks are Boosting techniques, neural networks and support vector machines (SVM) [16], [20].

IX. EXTENSIONS AND FUTURE WORK

Motivated by results in rate-distortion for lossy compression [6], [7], [49], one practical direction for addressing the MPE-SR problem is to sacrifice some approximation quality by restricting the problem to families of dictionaries with strong embedded structure, where by taking advantage of its hierarchical structure the operational R-D problem, Eq.(16), reduces to a combinatorial problem and can be addressed in polynomial time using DP techniques. We have followed this general direction and addressed the MPE-SR problem for a family of representations induced by the Wavelet packets (WP). This family provides rich time-frequency representations for the raw observation space and it has been used extensively for the analysis of pseudo-stationary random processes and also quasi-periodic random fields, with speech and image classification problems being some examples of that. Formulation of the MPE-SR problem in this scenario and some experimental results are presented in [58].

X. ACKNOWLEDGMENTS

The authors are grateful to Antonio Ortega for helpful discussion and Joseph Tepperman for proof reading this material. This material is based upon work supported by awards from National Science Foundation, ONR-MURI, DARPA and the U.S. Army.

APPENDIX I

PROOF OF THEOREM 1

Proof: Let us denote by $g(\cdot)$ and $\hat{g}(\cdot)$ the Bayes classification rules associated with $P_{X,Y}$ and $\hat{P}_{X,Y}$, respectively, Eq.(1). We can characterize the error probability for the decision $g(\cdot)$ as:

$$\begin{aligned}
\mathbb{P}(\{u \in \Omega : g(X(u)) \neq Y(u)\}) &= P_{X,Y}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : g(x) \neq y\}) \\
&= 1 - \mathbb{E}_{X,Y}(\mathbb{1}_{\{(x,y) \in \mathcal{X} \times \mathcal{Y} : g(x)=y\}}(X, Y)) \\
&= 1 - \mathbb{E}_Y(\mathbb{E}_{X|Y}(\mathbb{1}_{\{(x,y) \in \mathcal{X} \times \mathcal{Y} : g(x)=y\}}(X, Y)|Y)) \\
&= 1 - \sum_{y \in \mathcal{Y}} P_{X,Y}(\mathcal{X}_y \times \{y\})
\end{aligned} \tag{I.1}$$

where $\mathcal{X}_y \equiv g^{-1}(\{y\})$, $\forall y \in \mathcal{Y}$. The same is true for the decision rule $\hat{g}(\cdot)$:

$$P_{X,Y}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \hat{g}(x) \neq y\}) = 1 - \sum_{y \in \mathcal{Y}} P_{X,Y}(\bar{\mathcal{X}}_y \times \{y\}) \tag{I.2}$$

$\bar{\mathcal{X}}_y \equiv \hat{g}^{-1}(\{y\})$, $\forall y \in \mathcal{Y}$.

Consequently the performance degradations of using $\hat{g}(\cdot)$ instead of $g(\cdot)$ is given by:

$$\begin{aligned}
\Delta(\hat{g}|g) &\equiv P_{X,Y}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \hat{g}(x) \neq y\}) - P_{X,Y}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : g(x) \neq y\}) \\
&= \sum_{y \in \mathcal{Y}} P_{X,Y}(\mathcal{X}_y \times \{y\}) - P_{X,Y}(\bar{\mathcal{X}}_y \times \{y\})
\end{aligned} \tag{I.3}$$

Let us focus on one of the terms of this expression. Then we have that:

$$\begin{aligned}
P_{X,Y}(\mathcal{X}_y \times \{y\}) - P_{X,Y}(\bar{\mathcal{X}}_y \times \{y\}) &= P_{X,Y}(\mathcal{X}_y \setminus \bar{\mathcal{X}}_y \times \{y\}) - P_{X,Y}(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y \times \{y\}) \\
&= P_{X,Y}(\mathcal{X}_y \setminus \bar{\mathcal{X}}_y \times \{y\}) - \hat{P}_{X,Y}(\mathcal{X}_y \setminus \bar{\mathcal{X}}_y \times \{y\}) \\
&\quad + \hat{P}_{X,Y}(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y \times \{y\}) - P_{X,Y}(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y \times \{y\}) \\
&\quad + \hat{P}_{X,Y}(\mathcal{X}_y \setminus \bar{\mathcal{X}}_y \times \{y\}) - \hat{P}_{X,Y}(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y \times \{y\})
\end{aligned} \tag{I.4}$$

where we have that the two first terms in Eq(I.4) satisfy the following inequality:

$$P_{X,Y}(\mathcal{X}_y \setminus \bar{\mathcal{X}}_y \times \{y\}) - \hat{P}_{X,Y}(\mathcal{X}_y \setminus \bar{\mathcal{X}}_y \times \{y\}) + \hat{P}_{X,Y}(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y \times \{y\}) - P_{X,Y}(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y \times \{y\}) \tag{I.5}$$

$$\begin{aligned}
&\leq P_Y(\{y\}) \cdot \int_{(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y) \cup (\mathcal{X}_y \setminus \bar{\mathcal{X}}_y)} |p_{X|Y}(x|y) - \hat{p}_{X|Y}(x|y)| \partial x \\
&\leq P_Y(\{y\}) \cdot \int_{\mathcal{X}} |p_{X|Y}(x|y) - \hat{p}_{X|Y}(x|y)| \partial x
\end{aligned} \tag{I.6}$$

where $p_{X|Y}(\cdot|y)$ and $\hat{p}_{X|Y}(\cdot|y)$ are the pdfs of $P_{X|Y}(\cdot|y)$ and $\hat{P}_{X|Y}(\cdot|y)$, respectively. Using this last result in conjunction with Eqs.(I.3) and (I.4) we have that:

$$\begin{aligned}
\Delta(\hat{g}|g) &\leq \sum_{y \in \mathcal{Y}} P_Y(\{y\}) \cdot \int_{\mathcal{X}} |p_{X|Y}(x|y) - \hat{p}_{X|Y}(x|y)| \partial x \\
&\quad + \sum_{y \in \mathcal{Y}} \hat{P}_{X,Y}(\mathcal{X}_y \setminus \bar{\mathcal{X}}_y \times \{y\}) - \hat{P}_{X,Y}(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y \times \{y\})
\end{aligned} \tag{I.7}$$

where for the last term it follows that:

$$\begin{aligned}
& \sum_{y \in \mathcal{Y}} \hat{P}_{X,Y}(\mathcal{X}_y \setminus \bar{\mathcal{X}}_y \times \{y\}) - \hat{P}_{X,Y}(\bar{\mathcal{X}}_y \setminus \mathcal{X}_y \times \{y\}) \\
&= \sum_{y \in \mathcal{Y}} \hat{P}_{X,Y}(\mathcal{X}_y \times \{y\}) - \hat{P}_{X,Y}(\bar{\mathcal{X}}_y \times \{y\}) \\
&= \hat{P}_{X,Y}(\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \hat{g}(x) \neq y\}) - \hat{P}_{X,Y}(\{(x,y) \in \mathcal{X} \times \mathcal{Y} : g(x) \neq y\}) \leq 0
\end{aligned} \tag{I.8}$$

The last inequality is because $\hat{g}(\cdot)$ is the Bayes classification rule associated with $\hat{P}_{X,Y}(\cdot)$. Finally we have that:

$$\begin{aligned}
\Delta(\hat{g}||g) &\leq \sum_{y \in \mathcal{Y}} P_Y(\{y\}) \cdot \int_{\mathcal{X}} |p_{X|Y}(x|y) - \hat{p}_{X|Y}(x|y)| \partial x \\
&\leq \sum_{y \in \mathcal{Y}} P_Y(\{y\}) \cdot \sqrt{2 \log(2) D(P_{X|Y}(\cdot|y) || \hat{P}_{X|Y}(\cdot|y))}.
\end{aligned} \tag{I.9}$$

This last step uses Pinsker's inequality presented for the discrete case in [19] (*Lemma 12.6.1*). The final step is a consequence of the symmetry of the last inequality:

$$\Delta(\hat{g}||g) \leq \sqrt{2 \ln 2} \sum_{y \in \mathcal{Y}} P_Y(\{y\}) \cdot \sqrt{\min \left\{ D(P_{X|Y}(\cdot|y) || \hat{P}_{X|Y}(\cdot|y)), D(\hat{P}_{X|Y}(\cdot|y) || P_{X|Y}(\cdot|y)) \right\}}. \tag{I.10}$$

■

APPENDIX II

TRADEOFF BETWEEN BAYES ERROR AND ESTIMATION ERROR

Proof: For proving the Bayes bound inequality we invoke the result presented in LEMMA 1 about deterministic transformation of the observation space. Given that $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ is a sequence of embedded transformations, and consequently $\forall i \in \{1, \dots, n-1\}$, there exists a measurable mapping $\pi_{i+1,i} : \mathcal{X}_{i+1} \rightarrow \mathcal{X}_i$ such that $X_i(u) = \pi_{i+1,i}(X_{i+1}(u))$, then from the LEMMA 1 we have that

$$L_{\mathcal{X}_{i+1}} \leq L_{\mathcal{X}_i}. \tag{II.1}$$

For proving the inequality regarding the estimation error across the sequence of embedded spaces, a sufficient condition, given the result presented in THEOREM 1, is to prove that

$$D_{(\mathcal{X}_i, \mathcal{F}_i)}(P_{X_i|Y}(\cdot|y) || \hat{P}_{X_i|Y}(\cdot|y)) \leq D_{(\mathcal{X}_{i+1}, \mathcal{F}_{i+1})}(P_{X_{i+1}|Y}(\cdot|y) || \hat{P}_{X_{i+1}|Y}(\cdot|y)), \tag{II.2}$$

$$D_{(\mathcal{X}_i, \mathcal{F}_i)}(\hat{P}_{X_i|Y}(\cdot|y) || P_{X_i|Y}(\cdot|y)) \leq D_{(\mathcal{X}_{i+1}, \mathcal{F}_{i+1})}(\hat{P}_{X_{i+1}|Y}(\cdot|y) || P_{X_{i+1}|Y}(\cdot|y)), \tag{II.3}$$

$\forall i \in \{1, \dots, n-1\}$ and $\forall y \in \mathcal{Y}$. We will focus on proving Eq.(II.2), proving the other family of inequalities is equivalent.

In this equation we consider $D_{(\mathcal{X}_i, \mathcal{F}_i)}(P_{X_i|Y}(\cdot|y) || \hat{P}_{X_i|Y}(\cdot|y))$ as the Kullback-Leibler divergence (KLD) of the conditional class probability $P_{X_i|Y}(\cdot|y)$ with respect to the empirical counterpart in the measurable space $(\mathcal{X}_i, \mathcal{F}_i)$. The fact of considering the dependency with respect to the underlying measurable space in the KLD notation, which is usually implicit, is conceptually important for the rest of the proof.

The main idea is to represent the empirical distribution as an underlying measure defined on the original measurable space $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$. This is possible using the fact that functions in $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ are measurable. Consequently given the empirical class conditional probability $\hat{P}_{X_i|Y}(\cdot|y)$ in the representation space $(\mathcal{X}_i, \mathcal{F}_i)$, we can induce a probability measure $\hat{P}_{X|Y}(\cdot|y)$ in the measurable space $(\mathcal{X}, \sigma(\mathbb{F}_i))$, where $\sigma(\mathbb{F}_i)$ is the smallest sigma field that makes $\mathbb{F}_i(\cdot)$ a measurable transformation [34]. Given that $\mathbb{F}_i : (\mathcal{X}, \mathcal{F}_{\mathcal{X}}) \rightarrow (\mathcal{X}_i, \mathcal{F}_i)$ is measurable, we have that $\sigma(\mathbb{F}_i) \subset \mathcal{F}_{\mathcal{X}}$ [34]. More precisely, $\sigma(\mathbb{F}_i) = \{\mathbb{F}_i^{-1}(B) : B \in \mathcal{F}_{\mathcal{X}}\}$ and $\hat{P}_{X|Y}(\cdot|y)$ is constructed by

$$\forall A \in \sigma(\mathbb{F}_i), \exists B \in \mathcal{F}_i, \text{ st. } A = \mathbb{F}_i^{-1}(B) \text{ and } \hat{P}_{X|Y}(A|y) = \hat{P}_{X_i|Y}(B|y). \quad (\text{II.4})$$

By the consistence property of $\{\hat{P}_{X_i|Y}(\cdot|y) : i = 1, \dots, n\}$, it is easy to show that there is a unique measure $\hat{P}_{X|Y}(\cdot|y)$ defined on $(X, \sigma(\mathbb{F}_n))$ that represents the family of empirical distributions $\{\hat{P}_{X_i|Y}(\cdot|y) : i = 1, \dots, n\}$ using the procedure presented in Eq.(II.4). It is important to mention that this sequence of induced sigma fields characterizes a filtration [59], [60], in other words $\sigma(\mathbb{F}_i) \subset \sigma(\mathbb{F}_{i+1})$, because of the existence of a measurable mapping $\pi_{i+1,i}(\cdot)$ from $(\mathcal{X}_i, \mathcal{F}_i)$ to $(\mathcal{X}_{i+1}, \mathcal{F}_{i+1})$. As a consequence, the empirical measure $\hat{P}_{X|Y}(\cdot|y)$ is uniquely characterized in \mathcal{X} using the finest sigma field $\sigma(\mathbb{F}_n)$. On the other hand, the original probability measure $P_{X|Y}(\cdot|y)$ is originally defined on $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and given that $\sigma(\mathbb{F}_n) \subset \mathcal{F}_{\mathcal{X}}$, it extends naturally to $(X, \sigma(\mathbb{F}_i))$, $\forall i \in \{1, \dots, n\}$.

The next step is to represent the KLD in the representation space Eq.(II.2), as a KLD in the original observation space \mathcal{X} relative to a particular sigma field. Using a classical result from measure theory [34], it is possible to prove that [41](Lemma 5.2.4),

$$D_{(X, \sigma(\mathbb{F}_i))}(P_{X|Y}(\cdot|y) || \hat{P}_{X|Y}(\cdot|y)) = D_{(\mathcal{X}_i, \mathcal{F}_i)}(P_{X_i|Y}(\cdot|y) || \hat{P}_{X_i|Y}(\cdot|y)). \quad (\text{II.5})$$

Finally from proving our target inequality Eq.(II.2), we make use of the following lemma.

LEMMA 2: Let us consider two measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{X}, \bar{\mathcal{F}})$, such that $\bar{\mathcal{F}}$ is a refinement of \mathcal{F} , in other words $\mathcal{F} \subset \bar{\mathcal{F}}$. In addition, let us consider two probability measures P_1 and P_2 defined on $(\mathcal{X}, \bar{\mathcal{F}})$, then assuming that $P_1 \ll P_2$, the following inequality holds,

$$D_{(\mathcal{X}, \bar{\mathcal{F}})}(P_1 || P_2) \geq D_{(\mathcal{X}, \mathcal{F})}(P_1 || P_2). \quad (\text{II.6})$$

This lemma can be proved directly using the definition of KLD for standard spaces [41](Chapter 5.2), in particular \mathbb{R}^K . Details of the proof can be found in [41] (Lemma 5.2.5).

In our context we have $P_{X|Y}(\cdot|y)$ and $\hat{P}_{X|Y}(\cdot|y)$ defined on $(\mathcal{X}, \sigma(\mathbb{F}_{i+1}))$ and consequently on $(\mathcal{X}, \sigma(\mathbb{F}_i))$, because $\sigma(\mathbb{F}_{i+1})$ is a refinement of $\sigma(\mathbb{F}_i)$, then Eq.(II.2) follows directly from LEMMA 2. ■

APPENDIX III

TRADEOFF BETWEEN BAYES ERROR AND ESTIMATION ERROR: FINITE ALPHABET CASE

Proof: This proof follows the same argumentation as the one presented in Appendix II, but for the case of finite alphabet representation functions. Let us denote the Bayes classification rule for $(X_i(u), Y(u))$ by $g_{P_{X_i, Y}}(\cdot)$

with error probability $L_{\mathcal{A}_i}$, given by

$$L_{\mathcal{A}_i} = P_{X_i, Y}(\{(x, y) \in \mathcal{A}_i \times \mathcal{Y} : g_{P_{X_i, Y}}(x) \neq y\}), \quad (\text{III.1})$$

$\forall i \in \{1, \dots, n\}$.

By the assumption that the representation family $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ is embedded, then by *Definition 4*, we have that $\forall 0 \leq i < j \leq n$, $X_i(u) \equiv \mathbb{F}_i(X(u)) = \pi_{j,i}(\mathbb{F}_j(X(u))) = \pi_{j,i}(X_j(u))$. Consequently using *LEMMA 1*, the Bayes error bound inequality, $L_{\mathcal{A}_{i+1}} \leq L_{\mathcal{A}_i}$, $\forall i \in \{1, \dots, n-1\}$, follows directly.

For the estimation error inequality, we will prove the following sufficient condition:

$$D(P_{X_i|Y}(\cdot|y) || \hat{P}_{X_i|Y}(\cdot|y)) \leq D(P_{X_{i+1}|Y}(\cdot|y) || \hat{P}_{X_{i+1}|Y}(\cdot|y)), \quad (\text{III.2})$$

$$D(\hat{P}_{X_i|Y}(\cdot|y) || P_{X_i|Y}(\cdot|y)) \leq D(\hat{P}_{X_{i+1}|Y}(\cdot|y) || P_{X_{i+1}|Y}(\cdot|y)), \quad (\text{III.3})$$

$\forall i \in \{1, \dots, n-1\}$ and $\forall y \in \mathcal{Y}$. We focus on proving Eq.(III.2) because the other inequalities present the same derivations.

Without loss of generality we can consider one generic pair i_o, y_o . Let us consider the empirical distribution \hat{P}_{i_o} induced on $(\mathcal{X}, \sigma_{i_o})$ by the measurable transformation $\mathbb{F}_{i_o}(\cdot)$ and the probability space $(\mathcal{A}_i, \hat{P}_{X_{i_o}|Y}(\cdot|y_o))$. In this case σ_{i_o} is the sigma field induced by the partition $Q_{i_o} \equiv \{\mathbb{F}_{i_o}^{-1}(\{a\}) : a \in \mathcal{A}_{i_o}\}$, and consequently the measure \hat{P}_{i_o} is univocally characterized by

$$\hat{P}_{i_o}(\mathbb{F}_{i_o}^{-1}(\{a\})) = \hat{P}_{X_{i_o}|Y}(\{a\} | y_o), \quad (\text{III.4})$$

$\forall a \in \mathcal{A}_{i_o}$, [34].

The same process can be used to induce a measure \hat{P}_{i_o+1} on $(\mathcal{X}, \sigma_{i_o+1})$. Note that given that the family of representations is embedded, we have that Q_{i_o+1} is a refinement of Q_{i_o} in \mathcal{X} and consequently $\sigma_{i_o} \subset \sigma_{i_o+1}$, [34]. Then we have that \hat{P}_{i_o+1} is also well defined on $(\mathcal{X}, \sigma_{i_o})$.

Moreover, by the consistence property of the conditional class probabilities $\{\hat{P}_{X_i|Y}(\cdot|y_o) : i = 1, \dots, n\}$ on the family of representation functions $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$, we want to show that the two measures agree on σ_{i_o} . For that we just need to show that they agree on the set of events that generate the sigma field, i.e., in $Q_{i_o} = \{\mathbb{F}_{i_o}^{-1}(\{a\}) : a \in \mathcal{A}_{i_o}\}$, because Q_{i_o} is a partition and in particular a semi-algebra [34]. Then without loss of generality let us consider the event $\mathbb{F}_{i_o}^{-1}(\{a\})$, then we have that:

$$\begin{aligned} \hat{P}_{i_o+1}(\mathbb{F}_{i_o}^{-1}(\{a\})) &= \hat{P}_{i_o+1}(\mathbb{F}_{i_o+1}^{-1}(\pi_{i_o+1, i_o}^{-1}(\{a\}))) \\ &= \hat{P}_{X_{i_o+1}|Y}(\pi_{i_o+1, i_o}^{-1}(\{a\}) | y_o) \\ &= \hat{P}_{X_{i_o}|Y}(\{a\} | y_o) \\ &= \hat{P}_{i_o}(\mathbb{F}_{i_o}^{-1}(\{a\})) \end{aligned} \quad (\text{III.5})$$

$\forall a \in \mathcal{A}_{i_o}$.

The first equality is because of the fact that $\mathbb{F}_i(\cdot) = \pi_{i_o+1, i_o}(\mathbb{F}_j(\cdot))$ — embedded property of the representation family, the second by Eq.(III.4), the third by the consistence property of the conditional class probabilities and the last again by definition of $P_{i_o}(\cdot)$, Eq.(III.4).

Consequently, we can just consider $\bar{P} \equiv \bar{P}_{i_{o+1}}$ as the empirical probability measure well defined on $(\mathcal{X}, \sigma_{i_{o+1}})$ and $(\mathcal{X}, \sigma_{i_o})$. Also note that the original probability measure $P_{X|Y}(\cdot|y_o)$ is well defined on $(\mathcal{X}, \sigma_{i_{o+1}})$ and $(\mathcal{X}, \sigma_{i_o})$ by the measurability of \mathbb{F}_{i_o} and $\mathbb{F}_{i_{o+1}}$, respectively [34].

Finally, it is not difficult to prove using the definition of the divergence [41] (*Chapter 5*) that:

$$D(P_{X_{i_o}|Y}(\cdot|y_o)||\hat{P}_{X_{i_o}|Y}(\cdot|y_o)) = D_{(\mathcal{X}, \sigma_{i_o})}(P_{X|Y}(\cdot|y_o)||\hat{P}) \quad (\text{III.6})$$

$$D(P_{X_{i_{o+1}}|Y}(\cdot|y_o)||\hat{P}_{X_{i_{o+1}}|Y}(\cdot|y_o)) = D_{(\mathcal{X}, \sigma_{i_{o+1}})}(P_{X|Y}(\cdot|y_o)||\hat{P}) \quad (\text{III.7})$$

where,

$$D_{(\mathcal{X}, \sigma_{i_o})}(P_{X|Y}(\cdot|y_o)||\hat{P}) = \sum_{A \in \mathcal{Q}_{i_o}} P_{X|Y}(A|y_o) \log \frac{P_{X|Y}(A|y_o)}{\hat{P}(A)}$$

$$D_{(\mathcal{X}, \sigma_{i_{o+1}})}(P_{X|Y}(\cdot|y_o)||\hat{P}) = \sum_{A \in \mathcal{Q}_{i_{o+1}}} P_{X|Y}(A|y_o) \log \frac{P_{X|Y}(A|y_o)}{\hat{P}(A)}$$

and using the LEMMA 2 presented in *Appendix II*, considering that $\sigma_{i_o} \subset \sigma_{i_{o+1}}$, and Eqs. (III.6) and (III.7), we prove the sufficient condition stated in Eq.(III.2) and consequently the result. ■

APPENDIX IV

PROOF THAT MAXIMUM LIKELIHOOD ESTIMATION IS CONSISTENT WITH RESPECT TO A SEQUENCE OF FINITE EMBEDDED REPRESENTATIONS

Proof: Let us consider $\{\mathbb{F}_i(\cdot) : i = 1, \dots, n\}$ to be a family of embedded representation functions, *Definition 4*, taking values in finite alphabet sets $\{\mathcal{A}_i : i = 1, \dots, n\}$, respectively. Let us consider in addition a training set $\{(x_i, y_i) : i = 1, \dots, N\}$ of iid realizations of the random vector $(X(u), Y(u))$. For every representation space \mathcal{A}_i , the empirical distribution is obtained by the ML criterion [2], [14], where consequently the conditional class distribution is given by,

$$\hat{P}_{X_i|Y}(\{a\} | y) = \frac{\sum_{k=1}^N \mathbb{1}_{\{(a, y)\}}(\mathbb{F}_i(x_k), y_k)}{N_y}, \quad (\text{IV.1})$$

$\forall i \in \{1, \dots, n\}$, $\forall a \in \mathcal{A}_i$ and $\forall y \in \mathcal{Y}$, where $N_y \equiv \sum_{k=1}^N \mathbb{1}_{\{y\}}(y_k)$ is assumed to be strictly greater than zero.

For the proof we will use the induced probability measure on the original observation space $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$, that we define by

$$\hat{P}_{i|y}(\mathbb{F}_i^{-1}(\{a\})) \equiv \hat{P}_{X_i|Y}(\{a\} | y), \quad (\text{IV.2})$$

for all $i \in \{1, \dots, n\}$ and $y \in \mathcal{Y}$.

By Eq.(IV.1), it is straightforward to show that for any $a \in \mathcal{A}_i$

$$\hat{P}_{i|y}(\mathbb{F}_i^{-1}(\{a\})) = \frac{\sum_{k=1}^N \mathbb{1}_{\{(\mathbb{F}_i^{-1}(\{a\}), y)\}}(x_k, y_k)}{N_y}. \quad (\text{IV.3})$$

Without loss of generality, let us consider $i_o, j_o \in \{1, \dots, n\}$ and $y_o \in \mathcal{Y}$, such that $i_o < j_o$. For proving the consistence condition of the ML empirical distributions, we just need to show that

$$\hat{P}_{X_{i_o}|Y}(\{a\} | y_o) = \hat{P}_{X_{j_o}|Y}(\pi_{j_o, i_o}^{-1}(\{a\}) | y_o), \quad (\text{IV.4})$$

$\forall a \in \mathcal{A}_{io}$.

By *Remark 4*, we have that the induced quantization $Q_{F_{j_o}} \equiv \{\mathbb{F}_{j_o}^{-1}(\{a\}) : a \in \mathcal{A}_{j_o}\}$ is a refinement of $Q_{F_{i_o}} \equiv \{\mathbb{F}_{i_o}^{-1}(\{a\}) : a \in \mathcal{A}_{i_o}\}$. Then, any atom $\mathbb{F}_{j_o}^{-1}(\{a\})$ indexed by $a \in \mathcal{A}_{i_o}$, can be expressed as disjoint unions of atoms in $Q_{F_{j_o}}$; more precisely, we have that:

$$\begin{aligned} \mathbb{F}_{i_o}^{-1}(\{a\}) &= \bigcup_{b \in \pi_{j_o, i_o}^{-1}(\{a\})} \mathbb{F}_{j_o}^{-1}(\{b\}) \\ &= \mathbb{F}_{j_o}^{-1}(\pi_{j_o, i_o}^{-1}(\{a\})) \end{aligned} \quad (\text{IV.5})$$

where finally by Eqs. (IV.2) and (IV.3), we have that:

$$\begin{aligned} \hat{P}_{X_{i_o}|Y}(\{a\} | y_o) &= \frac{\sum_{k=1}^N \mathbb{1}_{\{\mathbb{F}_{i_o}^{-1}(a), y_o\}}(x_k, y_k)}{N_{y_o}} \\ &= \frac{\sum_{k=1}^N \mathbb{1}_{\{\mathbb{F}_{j_o}^{-1}(\pi_{j_o, i_o}^{-1}(\{a\})), y_o\}}(x_k, y_k)}{N_{y_o}} \\ &= \hat{P}_{X_{j_o}|Y}(\pi_{j_o, i_o}^{-1}(\{a\}) | y_o). \end{aligned} \quad (\text{IV.6})$$

■

APPENDIX V

PROOF THAT MAXIMUM LIKELIHOOD ESTIMATION IS CONSISTENT FOR THE GAUSSIAN PARAMETRIC ASSUMPTION

Proof: Without loss of generality, let us consider just $\mathbf{f}_1(x) = \mathbf{A}_1 \cdot x$ and $\mathbf{f}_2(x) = \mathbf{A}_2 \cdot x$, with $\mathbf{A}_1 \in \mathbb{R}(k1, K)$ and $\mathbf{A}_2 \in \mathbb{R}(k2, K)$ ($0 < k1 < k2 < K$). We need to show that $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot | y)$ defined on $(\mathbb{R}^{k2}, \mathcal{B}^{k2})$ is consistent with respect to $\hat{P}_{\mathbf{f}_1(X)|Y}(\cdot | y)$ defined on $(\mathbb{R}^{k1}, \mathcal{B}^{k1})$, in the sense that $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot | y)$ induces $\hat{P}_{\mathbf{f}_1(X)|Y}(\cdot | y)$ by the measurable mapping $B_{2,1} : (\mathbb{R}^{k2}, \mathcal{B}^{k2}) \rightarrow (\mathbb{R}^{k1}, \mathcal{B}^{k1})$. However under the Gaussian parametric assumption, this condition reduces to checking the first and second order statistics of the involved distributions. Considering the training data, it is direct to show that the empirical mean and covariance matrix for $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot | y)$ is given by $\mathbf{A}_2 \hat{\mu}_y$ and $\mathbf{A}_2 \hat{\Sigma}_y \mathbf{A}_2^\dagger$, respectively, where $\hat{\mu}_y$ and $\hat{\Sigma}_y$ are the respective empirical values in the original observation space \mathcal{X} , Eqs (35) and (36). Analogous results hold for the case of $\hat{P}_{\mathbf{f}_1(X)|Y}(\cdot | y)$.

Given that linear transformations preserve the multivariate Gaussian distribution, we have that $\hat{P}_{\mathbf{f}_2(X)|Y}(\cdot | y)$ induces a Gaussian distribution on $(\mathbb{R}^{k1}, \mathcal{B}^{k1})$ with mean $B_{2,1} \mathbf{A}_2 \hat{\mu}_y$ and covariance matrix $B_{2,1} \mathbf{A}_2 \hat{\Sigma}_y \mathbf{A}_2^\dagger B_{2,1}^\dagger$. Finally, given that the linear transformations $\mathbf{f}_1(\cdot)$ and $\mathbf{f}_2(\cdot)$ preserve the consistence structure of $\mathbb{R}^{k1}, \mathbb{R}^{k2}$, we have that $B_{2,1} \mathbf{A}_2 = \mathbf{A}_1$ which is sufficient to prove the result. ■

REFERENCES

- [1] N. Vasconcelos, "Minimum probability of error image retrieval," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2322–2336, August 2004.
- [2] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

- [3] V. K. Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in \mathbb{R}^n : Analysis, synthesis and algorithms," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 16–31, 1998.
- [4] Z. Cvetkovic and M. Vetterli, "On simple oversampled a/d conversion in $l_2(r)$," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 59–73, January 2001.
- [5] B. Beferull-Lozano and A. Ortega, "Efficient quantization for overcomplete expansions in \mathbb{R}^n ," *IEEE Transactions on Information Theory*, vol. 49, no. 1, pp. 129–150, January 2003.
- [6] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in rate-distortion sense," *IEEE Transactions on Image Processing*, vol. 2, no. 2, pp. 160–175, 1993.
- [7] K. Ramchandran, M. Vetterli, and C. Herley, "Wavelet, subband coding, and best bases," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 541–560, April 1996.
- [8] A. Cohen, I. Daubechies, O. Guleryuz, and M. Orchard, "On the importance of combining wavelet-based nonlinear approximation with coding strategies," *IEEE Transactions on Information Theory*, vol. 48, no. 1, July 2002.
- [9] S. Beheshti and M. A. Dahleh, "A new information-theoretic approach to signal denoising and best basis selection," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, October 2005.
- [10] M. Vetterli and J. Kovacevic, *Wavelet and Subband Coding*. Englewood Cliffs, NY: Prentice-Hall, 1995.
- [11] B. D. Rao and K. K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, January 1999.
- [12] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, August 2004.
- [13] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithm for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, March 1992.
- [14] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1983.
- [15] V. Vapnik, *Statistical Learning Theory*. John Wiley, 1998.
- [16] ———, *The Nature of Statistical Learning Theory*. Springer - Verlag, New York, 1999.
- [17] J. Wozencraft and I. Jacobs, *Principles of Communication Engineering*. Waveland Press, 1965.
- [18] S. Kullback, *Information theory and Statistics*. New York: Wiley, 1958.
- [19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Interscience, New York, 1991.
- [20] O. Bousquet, S. Boucheron, and G. Lugosi, *Theory of Classification: A Survey of Recent Advances*. ESAIM: Probability and Statistics, URL:<http://www.emath.fr/ps>, 2004.
- [21] M. Padmanabhan and S. Dharanipragada, "Maximizing information content in feature extraction," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 512–519, July 2005.
- [22] K. Etemad and R. Chellapa, "Separability-based multiscale basis selection and feature extraction for signal and image classification," *IEEE Transactions on Image Processing*, vol. 7, no. 10, pp. 1453–1465, October 1998.
- [23] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 6, pp. 2653–2667, November 1999.
- [24] S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1368–1379, July 2001.
- [25] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer Academic, 1992.
- [26] R. Gray, *Source Coding Theory*. Norwell, MA: Kluwer Academic, 1990.
- [27] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, pp. 2325–2384, October 1998.
- [28] T. F. Quatieri, *Discrete-time Speech Signal Processing principles and practice*. Prentice Hall, 2002.
- [29] J. Novovicova, P. Pudil, and J. Kittler, "Divergence based feature selection for multimodal class densities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, pp. 218–223, February 1996.
- [30] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4–36, April 2000.
- [31] T. M. Cover and J. M. V. Campenhout, "On the possible ordering in the measurement selection problem," *IEEE Transactions on Systems, Man, Cybern.*, vol. 7, pp. 657–661, 1977.

- [32] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2693–2723, 1998.
- [33] C. Scott, "Tree pruning with subadditive penalties," *IEEE Transactions on Signal Processing*, vol. 53, no. 12, pp. 4518–4525, 2005.
- [34] P. R. Halmos, *Measure Theory*. Van Nostrand, New York, 1950.
- [35] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
- [36] N. Vasconcelos, "Bayesian model for visual information retrieval," Ph.D. dissertation, Mass. Inst. of Technol., 2000.
- [37] N. A. Schmid and J. A. O'Sullivan, "Thresholding method for dimensionality reduction in recognition system," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2903–2920, November 2001.
- [38] G. F. Hughes, "On the mean accuracy of statistical pattern recognition," *IEEE Transactions on Information Theory*, vol. IT-14, no. 1, pp. 55–63, January 1968.
- [39] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–264, March 1991.
- [40] G. Trunk, "A problem of dimensionality: A simple example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 306–307, 1979.
- [41] R. M. Gray, *Entropy and Information Theory*. Springer - Verlag, New York, 1990.
- [42] C. E. Shannon, "Coding theorems for a discrete source with fidelity criterion," in *IRE Conv. Rec.*, vol. 7, no. 142-163, 1959.
- [43] A. B. Nobel, "Analysis of a complexity-based pruning scheme for classification tree," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2362–2368, 2002.
- [44] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423; 623–656, July and October 1948.
- [45] J. W. Fisher III, M. Wainwright, E. Sudderth, and A. S. Willsky, "Statistical and information-theoretic methods for self-organization and fusion of multimodal, networked," *International Journal of High Performance Computing Applications*, 2002.
- [46] J. W. Fisher III, T. Darrel, W. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing System, Denver, USA*. Advances in Neural Information Processing Systems, November 2000.
- [47] T. Butz and J.-P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Elsevier Signal Processing*, vol. 85, pp. 875–902, 2005.
- [48] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noise sources by means of waveform and autoregressive model vector quantization," *IEEE Transactions on Information Theory*, vol. 34, no. 4, pp. 826–834, 1998.
- [49] P. Chou, T. Lookabaugh, and R. Gray, "Optimal pruning with applications to tree-structure source coding and modeling," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 299–315, 1989.
- [50] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Transactions on Information Theory*, vol. it-16, no. 4, pp. 406–411, 1970.
- [51] H. S. Witsenhausen, "Indirect rate distortion problems," *IEEE Transactions on Information Theory*, vol. it-36, no. 5, pp. 518–521, 1980.
- [52] A. B. Nobel, "Recursive partitioning to reduce distortion," *IEEE Transactions on Information Theory*, vol. 43, no. 4, pp. 1122–1133, July 1997.
- [53] C. Scott and R. D. Nowak, "Minimax-optimal classification with dyadic decision trees," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1335–1353, April 2006.
- [54] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge Univ Press, 1996.
- [55] V. Vapnik, *Estimation of dependencies based on empirical Data*. Springer - Verlag, New York, 1979.
- [56] V. Vapnik and A. J. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability Apl.*, vol. 16, pp. 264–280, 1971.
- [57] J. Silva and S. Narayanan, "Upper bound kullback-leibler divergence for hidden markov models with application as discrimination measure for speech recognition," in *IEEE International Symposium on Information Theory*, July 2006.
- [58] —, "Optimal wavelet packets decomposition based on the minimum probability of error signal representation principle: Algorithms and applications," *submitted for review*, 2006.
- [59] L. Breiman, *Probability*. Addison-Wesley, 1968.
- [60] S. Varadhan, *Probability Theory*. American Mathematical Society, 2001.