

We choose

$$\Delta_1 = \Delta_2 + \frac{\phi_2 - \phi_1}{2} \in [-\phi_1, 0]. \quad (58)$$

A2) For the above choice of Δ_1

$$\phi_1 + 2\Delta_1 = \phi_2 + 2\Delta_2.$$

A1) Similarly

$$\begin{aligned} |\phi_1 + \Delta_1| &= \phi_1 + \Delta_1 \\ &= \frac{\phi_2}{2} + \frac{\phi_1}{2} + \Delta_2 \\ &\leq \phi_2 + \Delta_2 = |\phi_2 + \Delta_2|. \end{aligned}$$

A3) Also,

$$|\Delta_1| = -\Delta_1 = -\Delta_2 - \frac{\phi_2 - \phi_1}{2} \leq -\Delta_2 = |\Delta_2|.$$

III) Finally, let

$$\Delta_2 \in \left[-\frac{\phi_2 - \phi_1}{2}, 0 \right]. \quad (59)$$

A2) For (59)

$$\phi_2 + 2\Delta_2 \geq \phi_2 - (\phi_2 - \phi_1) = \phi_1.$$

Hence,

$$|\phi_2 + 2\Delta_2| \geq \phi_1 \geq |\phi_1 + 2\Delta_1|, \quad \forall \Delta_1 \in [-\phi_1, 0]. \quad (60)$$

A3) If $\Delta_2 < -\phi_1$ (which may happen if $\phi_2 > 3\phi_1$) then we choose $\Delta_1 = -\phi_1$; otherwise, we choose $\Delta_1 = \Delta_2$. Hence, by construction

$$|\Delta_2| \geq |\Delta_1|. \quad (61)$$

A1) As explained above, for $\Delta_2 < -\phi_1$ we choose $\Delta_1 = -\phi_1$ which implies that

$$|\phi_2 + \Delta_2| \geq |\phi_1 + \Delta_1| = 0.$$

For $\Delta_2 \geq -\phi_1$, we have $\Delta_1 = \Delta_2$ and, therefore,

$$|\phi_2 + \Delta_2| = \phi_2 + \Delta_2 = \phi_2 + \Delta_1 \geq \phi_1 + \Delta_1 = |\phi_1 + \Delta_1|.$$

The proof of (52) is therefore complete. The stated result in (32) follows easily from (33) and (52).

ACKNOWLEDGMENT

We would like to thank Dr. Marvin Simon of JPL for some useful comments on this correspondence.

REFERENCES

- [1] M. K. Simon and D. Divsalar, "On the implementation and performance of single and double differential detection schemes," *IEEE Trans. Commun.*, vol. 40, pp. 278–291, Feb. 1992.
- [2] D. K. van Alphen and W. C. Lindsey, "Higher-order differential phase shift keyed modulation," *IEEE Trans. Commun.*, vol. 42, pp. 440–448, Feb./Mar./Apr. 1994.
- [3] F. Gini and G. B. Giannakis, "Generalized differential encoding: A nonlinear signal processing framework," *IEEE Trans. Signal Processing*, vol. 46, pp. 2967–2974, Nov. 1998.
- [4] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*. Englewood Cliffs, NJ: Prentice-Hall PTR, 1994.
- [5] J. L. Buetefuer and W. G. Cowley, "Frequency offset insensitive multiple symbol detection of MPSK," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 5, 2000, pp. 2669–2672.
- [6] D. Divsalar and M. K. Simon, "Multiple-symbol differential detection of MPSK," *IEEE Trans. Commun.*, vol. 38, pp. 300–308, Mar. 1990.

A Note on Density Model Size Testing

G rard Biau and Luc Devroye

Abstract—Let $(\mathcal{F}_k)_{k \geq 1}$ be a nested family of parametric classes of densities with finite Vapnik–Chervonenkis dimension. Let f be a probability density belonging to $\mathcal{F}_{k^*}$, where k^* is the unknown smallest integer such that $f \in \mathcal{F}_k$. Given a random sample $X_1, \dots, X_n$ drawn from f , an integer $k_0 \geq 1$ and a real number $\alpha \in (0, 1)$, we introduce a new, simple, explicit α -level consistent testing procedure of the null hypothesis $\{\mathbf{H}_0 : k^* = k_0\}$ versus the alternative $\{\mathbf{H}_1 : k^* \neq k_0\}$. Our method is inspired by the combinatorial tools developed in Devroye and Lugosi [1] and it includes a wide range of density models, such as mixture models, neural networks, or exponential families.

Index Terms—Hypothesis testing, mixture densities, nonparametric estimation, penalization, Vapnik–Chervonenkis dimension.

I. INTRODUCTION

Let f be an unknown density on $\mathbb{R}^d$ belonging to a prespecified class of densities, $\mathcal{F}_k$, where k is unknown, but $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k . Define

$$\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k.$$

In the union above, $\mathcal{F}_k$ denotes, for each fixed $k \geq 1$, a given class of densities—often parameterized by one or more parameters—and considered from a topological point of view as a closed metric subspace of the space of all densities on $\mathbb{R}^d$ endowed with the L_1 metric. Note that the requirement that $\mathcal{F}_k$ is closed for the L_1 metric is not restrictive, since any metric subspace of L_1 can be extended into a closed one by the principle of extension by continuity (Dunford and Schwartz [2]). For example, $\mathcal{F}_k$ might be the class of all mixtures of k Gaussians on $\mathbb{R}^d$. Since $f \in \mathcal{F}$, it is natural to define the index of the economical representation of f (James, Priebe, and Marchette [3]) as

$$k^* = \min\{k \geq 1 : f \in \mathcal{F}_k\}.$$

Naturally, as it is assumed that $f \in \mathcal{F}$, one has $k^* < \infty$. Roughly speaking, k^* represents the index of the most parsimonious model for f . Now, given a random sample $X_1, \dots, X_n$ drawn from f , an integer $k_0 \geq 1$ and a real number $\alpha \in (0, 1)$, the purpose of this correspondence is to introduce a new, simple, explicit α -level consistent testing procedure of the null hypothesis $\{\mathbf{H}_0 : k^* = k_0\}$ versus the alternative $\{\mathbf{H}_1 : k^* \neq k_0\}$. The problem of testing hypotheses on k^* has received some attention in the literature, essentially for mixture classes, such as mixtures of Gaussians. Due to lack of identifiability, the limiting distribution for the likelihood ratio test statistic was until recently unavailable, and so bootstrap testing methodologies have been developed (see, e.g., McLachlan [4]). This lack of identifiability leads to the degeneracy of the Fisher information of the model, so that the classical chi-square theory does not apply. Dacunha-Castelle and Gassiat [5], [6] proposed a theory of reparametrization to solve such problems, that they called "locally conic parametrization." Roughly speaking, the idea is to approach the null hypothesis using directional submodels in

Manuscript received December 10, 2002; revised July 21, 2003.

G. Biau is with the Laboratoire de Statistique Th orique et Appliqu e, Universit  Pierre et Marie Curie–Paris VI, 75013 Paris, France (e-mail: biau@ccr.jussieu.fr).

L. Devroye is with the School of Computer Science, McGill University, Montreal, QC H3A 2K6, Canada (e-mail: luc@cs.mcgill.ca).

Communicated by A. B. Nobel, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Digital Object Identifier 10.1109/TIT.2004.825250

which the Fisher information is normalized to be uniformly equal to one.

In the present manuscript, we propose a new testing methodology, which is close in spirit to Biau and Devroye [7], where we show how to pick automatically, and without extra restrictions on f , a density estimate $f_{n,\hat{k}}$ in $\mathcal{F}$ with the property that

$$\mathbf{E} \left\{ \int |f_{n,\hat{k}} - f| \right\} = O \left(\frac{1}{\sqrt{n}} \right).$$

Our approach is inspired by the combinatorial tools developed in Devroye and Lugosi [1] and recently extended to the problem of robust hypothesis testing in Devroye, Györfi, and Lugosi [8]. As we explain it in [7], our methodology is not proper to mixtures, and it includes a wide range of models, such as neural networks or exponential families. We refer the reader to [7] for a detailed discussion, examples, and references. In fact, by taking the more classical statistical view, and concentrating on identification of the parameters, one is doomed to run into problems of identifiability and unstable or nonconverging estimation algorithms. Rather than focusing on the parameters, we will look directly at the performance of the estimate without worrying about the consistency in the space of all unknown parameters.

The correspondence is organized as follows. In Section II, we present our testing procedure as well as some useful related tools. The main results, level and consistency of the testing methodology, are stated in Section III. Proofs are gathered in Section IV.

II. THE TESTING PROCEDURE

A. Presentation

In [1], Devroye and Lugosi explore a new paradigm for the data-based or automatic selection of the free parameters of density estimates in general, so that the expected error is within a given constant multiple of the best possible error. To summarize in the present context, fix $k \geq 1$, and define a density estimate $f_{n,k}$ in $\mathcal{F}_k$ as follows. First introduce the class of sets

$$\mathcal{A}_k = \left\{ \{x : f(x) \geq g(x)\} : f, g \in \mathcal{F}_k \right\}$$

($\mathcal{A}_k$ is the so-called *Yatracos class* associated with $\mathcal{F}_k$) and the goodness criterion for a density $g \in \mathcal{F}_k$

$$\Delta_k(g) = \sup_{A \in \mathcal{A}_k} \left| \int_A g - \mu_n(A) \right|,$$

where

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{[X_i \in A]}$$

is the empirical measure associated with the sample $X_1, \dots, X_n$. For each $k \geq 1$, the minimum-distance estimate $f_{n,k}$ is defined as any density estimate selected from among those densities $f \in \mathcal{F}_k$ with

$$\Delta_k(f) < \inf_{g \in \mathcal{F}_k} \Delta_k(g) + \frac{1}{n}.$$

Note that the $1/n$ term here is added to ensure the existence of such a density estimate.

Having disposed of this preliminary step, we let $\mathcal{V}_k$ be the *Vapnik–Chervonenkis dimension* of the class of sets $\mathcal{A}_k$ (Vapnik and Chervonenkis [9]). Recall that $\mathcal{V}_k$ is defined as the largest integer p such that

$$\mathcal{S}_{\mathcal{A}_k}(p) = 2^p$$

where $\mathcal{S}_{\mathcal{A}_k}(p)$ is the Vapnik–Chervonenkis shatter coefficient, defined by

$$\mathcal{S}_{\mathcal{A}_k}(p) = \max_{x_1, \dots, x_p \in \mathbb{R}^d} \text{Card} \left\{ \{x_1, \dots, x_p\} \cap A : A \in \mathcal{A}_k \right\}.$$

If $\mathcal{S}_{\mathcal{A}_k}(p) = 2^p$ for all p , then we say that $V = \infty$. Using combinatorial arguments, one can show (see Devroye and Lugosi [1, Ch. 4] and the references therein) that if $\mathcal{A}_k$ has Vapnik–Chervonenkis dimension $\mathcal{V}_k$, then

$$\mathbf{E} \left\{ \Delta_k(f) \right\} \leq C \sqrt{\frac{\mathcal{V}_k}{n}} \quad (1)$$

where C is a universal positive constant. Now, denoting by $\alpha \in (0, 1)$ the required testing level, the proposed testing procedure is as follows: accept the null hypothesis $\mathbf{H}_0$ if

$$k_0 \in \arg \min_{k \geq 1} \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) \right\}. \quad (2)$$

Here, the term $\text{pen}_n(k)$ denotes a *penalty function* defined by

$$\text{pen}_n(k) = \begin{cases} -n^{-1/3} - 1/n, & \text{for } k = 1, \dots, k_0 - 1 \\ 0, & \text{for } k = k_0 \\ \frac{x_k + C \sqrt{\mathcal{V}_{k_0}}}{\sqrt{n}} + 1/n, & \text{for } k \geq k_0 + 1 \end{cases}$$

where the letter C stands for the universal constant of inequality (1) and $(x_k)_{k \geq k_0+1}$ is a sequence of nonnegative real numbers satisfying the condition

$$2 \sum_{k \geq k_0+1} e^{-2x_k^2} \leq \alpha/2.$$

Note that $\text{pen}_n(k)$ tends to infinity with k . This, together with the fact that the right-hand term of (2) is at most $1 + \text{pen}_n(1)$, shows that we need only do the computations for those k for which $\text{pen}_n(k) \leq 1 + \text{pen}_n(1)$. Therefore, a good choice for the sequence $(x_k)_{k \geq k_0+1}$ is one for which this sequence has a fast rate of divergence. For example, the choice

$$2e^{-2x_k^2} = \frac{\alpha}{2^{k+1}}$$

is preferable to the choice

$$2e^{-2x_k^2} = \frac{\alpha}{2k(k+1)}.$$

Note, however, that there is a price of letting x_k grow rapidly, see Remark 2 in Section III.

B. Some Useful Results

For each minimum-distance estimate $f_{n,k}$, we have (Devroye and Lugosi [1, Theorem 6.4])

$$\int |f_{n,k} - f| \leq 3 \inf_{g \in \mathcal{F}_k} \int |f - g| + 4\Delta_k(f) + \frac{3}{n}. \quad (3)$$

Since $f \in \mathcal{F}_{k^*}$, this inequality reduces to

$$\int |f_{n,k} - f| \leq 4\Delta_k(f) + \frac{3}{n}$$

as soon as $k \geq k^*$. Note that if not all members of $\mathcal{F}_k$ are densities (just take for $\mathcal{F}_k$ the class of all neural networks with k hidden nodes or the class of series estimates with k basis functions) then, as shown in Devroye and Lugosi [1, Exercise 6.2], (3) can be replaced by

$$\int |f_{n,k} - f| \leq 5 \inf_{g \in \mathcal{F}_k} \int |f - g| + 4\Delta_k(f) + \frac{5}{n}.$$

The only impact of this is that $3/n$ has to be replaced by $5/n$. The effect is so slight that the reader can easily carry through the necessary adjustments in the sections that follow. Thus, we will assume in the sequel that all members in all classes $\mathcal{F}_k$ are densities.

With respect to the term $\Delta_k(f)$ in (3), a simple consequence of the well-known bounded difference inequality (McDiarmid [10]) tells us that

$$P\left\{\left|\Delta_k(f) - \mathbf{E}\{\Delta_k(f)\}\right| > t\right\} \leq 2e^{-2nt^2} \quad (4)$$

for any $n \geq 1$ and $t > 0$. This shows that for any class $\mathcal{A}_k$, the maximal deviation is sharply concentrated around its mean. Combining (1) and (4) leads to the following useful inequality:

$$P\left\{\Delta_k(f) > \frac{t}{\sqrt{n}} + C\sqrt{\frac{\mathcal{V}_k}{n}}\right\} \leq 2e^{-2t^2}. \quad (5)$$

Finally, we will also use the following result, due to Talagrand [11], which states that there exists a universal positive constant D such that for all $n \geq 1$ and $\varepsilon > 0$

$$P\left\{\Delta_k(f) > \varepsilon\right\} \leq \frac{D}{\sqrt{n\varepsilon}} \left(\frac{Dn\varepsilon^2}{\mathcal{V}_k}\right)^{\mathcal{V}_k} e^{-2n\varepsilon^2} \quad (6)$$

whenever $\mathcal{V}_k$ is finite.

III. RESULTS

Here and in what follows, $\mathcal{B}$ denotes the class of all Borel sets of $\mathbb{R}^d$. Recall that a class $\mathcal{A}$ of subsets of $\mathcal{B}$ is a π -system if it is closed under the formation of finite intersections: $A, B \in \mathcal{A}$ implies $A \cap B \in \mathcal{A}$. See Billingsley [12] for details. We remind the reader that k^* , the index of the economical representation of f , is defined by

$$k^* = \min\{k \geq 1 : f \in \mathcal{F}_k\}.$$

Our first result insures that the testing procedure defined in Section II is of level α .

Theorem 1: Let $(\mathcal{A}_k)_{k \geq 1}$ be the sequence of Yatacos classes associated with the models $(\mathcal{F}_k)_{k \geq 1}$. Let $k_0 \geq 1$ be an integer and α a real number in $(0, 1)$. Assume that the Vapnik–Chervonenkis dimension $\mathcal{V}_{k_0}$ of $\mathcal{A}_{k_0}$ is finite. Then, provided $\mathcal{A}_1$ contains a π -system that generates $\mathcal{B}$, the testing procedure of $\{\mathbf{H}_0 : k^* = k_0\}$ versus $\{\mathbf{H}_1 : k^* \neq k_0\}$ as defined in (2) satisfies, for all n large enough

$$P_0\{\text{rejecting } \mathbf{H}_0\} \leq \alpha$$

where P_0 stands for the probability under the null hypothesis.

Remark 1: Theorem 1 holds for all n large enough, depending on the unknown f . From the proof it is easy to get estimates for the value of n necessary for the theorem to hold in terms of how well f can be approximated by densities in $\mathcal{F}_{k_0}$.

The next theorem deals with the consistency of the proposed test.

Theorem 2: Under the notations and assumptions of Theorem 1, with the additional condition that $\mathcal{V}_{k^*} < \infty$, there exists a positive constant K (depending on f) such that, for all $n \geq 1$

$$P_1\{\text{rejecting } \mathbf{H}_0\} \geq 1 - Kn^{-1/2 + \mathcal{V}_{k^*}} e^{-2n^{1/3}}$$

where P_1 stands for the probability under the alternative hypothesis.

Remark 2: It is seen from the proof that the value of the constant K increases with the value of the particular weight x_{k^*} whenever $k^* > k_0$. Since k^* is unknown, a good choice for the sequence $(x_k)_{k \geq k_0+1}$ is thus one for which this sequence has a slow rate of divergence. This, together with the discussion at the end of Section II-A, shows that the sequence $(x_k)_{k \geq k_0+1}$ should increase *reasonably fast* toward infinity.

As an illustration, just consider the classes $\mathcal{F}_k$ of all mixtures of k normal densities over $\mathbb{R}$, that is, the classes of all densities of form

$$f(x) = \sum_{i=1}^k \frac{p_i}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}(x-m_i)^2/\sigma_i^2} \quad (7)$$

where $(p_1, \dots, p_k)$ is a probability vector, $\sigma_1, \dots, \sigma_k$ are positive real numbers, and $m_1, \dots, m_k$ are arbitrary elements of $\mathbb{R}$. An enormous body of literature exists regarding the application, computational issues, and theoretical aspects of mixture models when the number of components is known, but estimating and testing the unknown number of components remains an area of intense research. The scope of application is vast, as mixture models are routinely employed across the entire diverse application range of statistics, including nearly all of the social and experimental sciences. For early references, see Everitt and Hand [13], Titterton, Smith, and Makov [14], McLachlan and Basford [15], and McLachlan and Peel [16]. The commonly used method for estimating the parameters of a mixture is the expectation–maximization (EM) algorithm (see Redner and Walker [17]). While originally designed for fixed mixture classes, such as mixtures of k Gaussians, the problem of the unknown k has received some attention in the Bayesian literature (Diebolt and Robert [18], Richardson and Green [19], Roeder and Wasserman [20], Celeux, Hurn, and Robert [21], and Hurn, Justel, and Robert [22]). The statistical learning community has also looked in depth at the problem (Bishop [23], Jordan and Jacobs [24], Zeevi and Meir [25], Figueiredo and Jain [26]). In clustering, or unsupervised learning, one often makes an assumption about the number of clusters and the distribution within each cluster. Estimating the distributions in the clusters and the weights of the clusters then leads to a natural way of clustering. Likelihood ratios have been used for this in most works, from Hartigan [27] to Fukumizu [28]. Dacunha-Castelle and Gassiat [5], [29], [6], on the other hand, use the moment method for identification and estimation of the number of components. The most recent attempts at estimating the mixture density parameters and the number of mixture densities jointly are by Priebe [30], James, Priebe, and Marchette [3], and Rogers, Marchette, and Priebe [31].

We draw attention on the fact that the conditions required in Theorems 1 and 2 are in no way restrictive and are in particular satisfied by a large choice of models. The requirement that $\mathcal{A}_1$ —and thus each $\mathcal{A}_k$ —contains a π -system generating Borel sets $\mathcal{B}$ is essentially of technical nature and in no way restrictive. Observe, for example, that it is satisfied for $d = 1$ as soon as the Yatacos class contains the π -system of all intervals, and more generally for $d \geq 1$ the π -system of all d -dimensional rectangles. In particular, it is satisfied by example (7). Moreover, it can be shown (see Devroye and Lugosi [1, Ch. 8]) that $\mathcal{V}_k = O(k^d)$, which ensures here that $\mathcal{V}_k < \infty$ for all $k \geq 1$. For more details and examples, we refer the reader to Biau and Devroye [7], where we use a close penalized combinatorial criterion to automatically pick a mixture complexity and a density from the given mixture, and still guarantee an $O(1/\sqrt{n})$ rate of convergence for the expected L_1 error, just as if we had been given the mixture complexity beforehand. Observe, however, that the role played by the penalty in the present testing problem is slightly different from the role played by the penalty in the density selection problem studied in [7]. Roughly speaking, the penalty function allows here to control the level of the testing procedure, whereas in [7], it guarantees a good rate of convergence by limiting the number of selected mixture components.

Note that the idea of using the additional x_k 's in the definition of the penalty is due to Barron, Birgé, and Massart [32], who studied performance bounds for model selection based on an empirical loss or contrast function with an added penalty term motivated by empirical process theory, and roughly proportional to the number of parameters needed to describe the model divided by the number of observations.

See also Castellan [33] and Massart [34]. Rissanen [35], [36] and Rissanen, Speed, and Yu [37] proposed model selection based on minimal stochastic complexity or description length. The focus there was principally on rate of convergence. It is not directly obvious how to modify these methods for hypothesis testing.

Finally, we draw attention on the fact that we can surely use the test for testing $k \leq k_0$ versus $k > k_0$. Just apply the test k_0 times for $k = \ell$ versus $k \neq \ell$, $\ell = 1, \dots, k_0$. If it is successful in one of these k_0 tests, then we say that $k \leq k_0$ is successful. The details can easily be worked out by the reader.

IV. PROOFS

A. Proof of Theorem 1

Before proving Theorem 1, we state a technical proposition.

Proposition 1: Let $k \geq 1$ and $\mathcal{A}_k$ be the Yatracos class associated with $\mathcal{F}_k$. Assume that $\mathcal{F}_k$ is closed for the L_1 metric on densities and that $\mathcal{A}_k$ contains a π -system that generates $\mathcal{B}$. Then $\mathcal{F}_k$ is closed for the D_k metric on densities, defined by

$$D_k(f, g) = \sup_{A \in \mathcal{A}_k} \left| \int_A f - \int_A g \right|.$$

Proof: First observe that the fact that $\mathcal{A}_k$ contains a π -system generating $\mathcal{B}$ forces D_k to be a metric on densities (see, for example, Billingsley [12]). Note also that, according to Scheffé's identity (Devroye [38, p. 2]), for two densities f and g

$$D_k(f, g) \leq \frac{1}{2} \int |f - g| \quad (8)$$

and, whenever $f, g \in \mathcal{F}_k$

$$D_k(f, g) = \frac{1}{2} \int |f - g|. \quad (9)$$

Now, let $(f_n)_{n \geq 1}$ be a Cauchy sequence in $\mathcal{F}_k$ for the D_k metric. Clearly, according to (9), $(f_n)_{n \geq 1}$ is also a Cauchy sequence for the L_1 metric. By assumption, $\mathcal{F}_k$ is closed for the L_1 metric on densities. Since the subspace of densities is closed in the complete space L_1 , $\mathcal{F}_k$ is also complete as a subspace of densities. Therefore, one deduces that there exists $f \in \mathcal{F}_k$ such that $\int |f_n - f| \rightarrow 0$ as $n \rightarrow \infty$. According to (8), this implies that $D_k(f_n, f) \rightarrow 0$ as $n \rightarrow \infty$. Thus, the set $\mathcal{F}_k$ is complete, and therefore closed, for the D_k metric. $\square$

Proof of Theorem 1: The following chain of inequalities is valid.

$$\begin{aligned} & \mathbf{P}_0 \{\text{rejecting } \mathbf{H}_0\} \\ &= \mathbf{P}_0 \left\{ \min_{k \neq k_0} \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) \right\} \right. \\ & \quad \left. < \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| \right\} \\ &\leq \sum_{\substack{k \geq 1 \\ k \neq k_0}} \mathbf{P}_0 \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) \right. \\ & \quad \left. < \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| \right\} \\ &\leq \sum_{\substack{k \geq 1 \\ k \neq k_0}} \mathbf{P}_0 \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) - \frac{1}{n} \right. \\ & \quad \left. < \Delta_{k_0}(f) \right\}, \\ & \quad (\text{since, under } \mathbf{H}_0, f \in \mathcal{F}_{k_0} \text{ and by definition of } f_{n,k_0}) \end{aligned}$$

where we recall that

$$\Delta_{k_0}(f) = \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f - \mu_n(A) \right|.$$

Therefore, we obtain that

$$\begin{aligned} & \mathbf{P}_0 \{\text{rejecting } \mathbf{H}_0\} \\ &\leq \sum_{\substack{k \geq 1 \\ k \neq k_0}} \mathbf{P}_0 \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) - \frac{1}{n} \right. \\ & \quad \left. < \Delta_{k_0}(f) \right\} \\ &= \sum_{k=1}^{k_0-1} \mathbf{P}_0 \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) - \frac{1}{n} \right. \\ & \quad \left. < \Delta_{k_0}(f) \right\} \\ & \quad + \sum_{k \geq k_0+1} \mathbf{P}_0 \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) - \frac{1}{n} \right. \\ & \quad \left. < \Delta_{k_0}(f) \right\}. \quad (10) \end{aligned}$$

We shall first examine the first of the two terms in (10) (which makes sense only if $k_0 > 1$). For $k = 1, \dots, k_0 - 1$, write

$$\begin{aligned} & \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| \\ &\geq \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \int_A f \right| - \sup_{A \in \mathcal{A}_k} \left| \int_A f - \mu_n(A) \right| \\ & \quad (\text{by the triangle inequality}) \\ &\geq \min_{1 \leq k \leq k_0-1} \inf_{g \in \mathcal{F}_k} \sup_{A \in \mathcal{A}_k} \left| \int_A f - \int_A g \right| - \Delta_{k_0}(f) \\ & \quad (\text{since } \mathcal{A}_k \subset \mathcal{A}_{k_0}) \\ &:= m - \Delta_{k_0}(f). \end{aligned}$$

By assumption, $\mathcal{A}_1$ —and thus each $\mathcal{A}_k$ —contains a π -system generating $\mathcal{B}$. Since the $\mathcal{F}_k$'s are closed for the L_1 metric on densities, we know from Proposition 1 that they are also closed for the D_k metric. Therefore, the definition of k_0 (recall that $k^* = k_0$ under $\mathbf{H}_0$) implies $m > 0$. Thus, we are led to

$$\begin{aligned} & \sum_{k=1}^{k_0-1} \mathbf{P}_0 \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f) \right\} \\ &\leq \sum_{k=1}^{k_0-1} \mathbf{P}_0 \left\{ 2\Delta_{k_0}(f) > m + \text{pen}_n(k) - \frac{1}{n} \right\} \\ &\leq (k_0 - 1) \mathbf{P}_0 \left\{ \Delta_{k_0}(f) > \frac{m}{4} \right\} \text{ for all } n \text{ large enough} \\ &\leq \alpha/2 \text{ for all } n \text{ large enough} \end{aligned}$$

where, in the last inequality, we used the finiteness of $\mathcal{V}_{k_0}$ together with inequality (5).

Let us now turn to the analysis of the second term in (10). We have

$$\begin{aligned} & \sum_{k \geq k_0+1} \mathbf{P}_0 \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) - \frac{1}{n} \right. \\ & \quad \left. < \Delta_{k_0}(f) \right\} \\ &\leq \sum_{k \geq k_0+1} \mathbf{P}_0 \left\{ \text{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f) \right\} \end{aligned}$$

$$\begin{aligned}
&= \sum_{k \geq k_0+1} \mathbf{P}_0 \left\{ \Delta_{k_0}(f) > \frac{x_k + C\sqrt{\mathcal{V}_{k_0}}}{\sqrt{n}} \right\} \\
&\quad (\text{by definition of the penalty function for } k \geq k_0 + 1) \\
&\leq 2 \sum_{k \geq k_0+1} e^{-2x_k^2} \quad (\text{by inequality (5)}) \\
&\leq \alpha/2 \quad (\text{by definition of the } x_k \text{'s}).
\end{aligned}$$

Putting all pieces together leads to the desired result.

B. Proof of Theorem 2

To prove Theorem 2, we show that $\mathbf{P}_1\{\text{accepting } \mathbf{H}_0\}$ goes to 0 as n grows. We have

$$\begin{aligned}
&\mathbf{P}_1\{\text{accepting } \mathbf{H}_0\} \\
&= \mathbf{P}_1 \left\{ \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| \right. \\
&\quad \left. = \min_{k \geq 1} \left\{ \sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \text{pen}_n(k) \right\} \right\} \\
&\leq \mathbf{P}_1 \left\{ \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| \right. \\
&\quad \left. \leq \sup_{A \in \mathcal{A}_{k^*}} \left| \int_A f_{n,k^*} - \mu_n(A) \right| + \text{pen}_n(k^*) \right\} \\
&\leq \mathbf{P}_1 \left\{ \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| - \text{pen}_n(k^*) - \frac{1}{n} \right. \\
&\quad \left. < \Delta_{k^*}(f) \right\}
\end{aligned}$$

where, in the last inequality, we use the definition of f_{n,k^*} and the fact that $f \in \mathcal{F}_{k^*}$.

Again, we distinguish the case $k_0 < k^*$ from the case $k_0 > k^*$ (recall that $k^* \neq k_0$ under $\mathbf{H}_1$). In the first situation (which makes sense only if $k^* > 1$), we have, acting as in the proof of Theorem 1

$$\begin{aligned}
\sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| &\geq m - \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f - \mu_n(A) \right| \\
&\geq m - \Delta_{k^*}(f)
\end{aligned}$$

where m is a positive constant. Therefore,

$$\mathbf{P}_1\{\text{accepting } \mathbf{H}_0\} \leq \mathbf{P}_1 \left\{ \Delta_{k^*}(f) > \frac{m}{4} \right\}$$

for all n large enough (depending on f), and this last term is bounded above by $K_1 n^{-1/2 + \mathcal{V}_{k^*}} e^{-nm^2/8}$, according to Talagrand's inequality (6), where K_1 is a positive constant depending on k^* . Finally, if k_0 meets the condition $k_0 > k^*$ (which makes sense only if $k_0 > 1$), then

$$\begin{aligned}
\mathbf{P}_1\{\text{accepting } \mathbf{H}_0\} &\leq \mathbf{P}_1 \left\{ -\text{pen}_n(k^*) - \frac{1}{n} < \Delta_{k^*}(f) \right\} \\
&\leq \mathbf{P}_1 \left\{ \Delta_{k^*}(f) > \frac{1}{n^{1/3}} \right\}
\end{aligned}$$

by definition of the penalty function for $k = 1, \dots, k_0 - 1$. We deduce again from Talagrand's inequality that the last term is bounded above by $K_2 n^{-1/6 + \mathcal{V}_{k^*}/3} e^{-2n^{1/3}}$ for all $n \geq 1$, where K_2 is a positive constant depending on k^* .

ACKNOWLEDGMENT

The authors thank both referees for their comments and suggestions.

REFERENCES

- [1] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer-Verlag, 2001.
- [2] N. Dunford and J. T. Schwartz, *Linear Operators Part I*. New York: Wiley, 1963.
- [3] L. F. James, C. E. Priebe, and D. J. Marchette, "Consistent estimation of mixture complexity," *Ann. Statist.*, vol. 29, pp. 1281–1296, 2001.
- [4] G. J. McLachlan, "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *J. Appl. Statist.*, vol. 36, pp. 318–324, 1987.
- [5] D. Dacunha-Castelle and E. Gassiat, "Testing in locally conic models, and application to mixture models," *ESAIM Probab. Statist.*, vol. 1, pp. 285–317, 1997.
- [6] —, "Testing the order of a model using locally conic parametrization: Population mixtures and stationary arma processes," *Ann. Statist.*, vol. 27, pp. 1178–1209, 1999.
- [7] G. Biau and L. Devroye, "Density estimation by the penalized combinatorial method," McGill University, Montreal, QC, Canada, Tech. Rep., 2002.
- [8] L. Devroye, L. Györfi, and G. Lugosi, "A note on robust hypothesis testing," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2111–2114, July 2002.
- [9] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Applic.*, vol. 16, pp. 264–280, 1971.
- [10] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics 1989*. Cambridge, U.K.: Cambridge Univ. Press, 1989.
- [11] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *Ann. Probab.*, vol. 22, pp. 28–76, 1994.
- [12] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.
- [13] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*. London, U.K.: Chapman and Hall, 1981.
- [14] D. M. Titterton, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. Chichester, U.K.: Wiley, 1985.
- [15] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.
- [16] G. J. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.
- [17] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, pp. 195–239, 1984.
- [18] J. Diebolt and C. P. Robert, "Estimation of finite mixture distributions through bayesian sampling," *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, vol. 56, pp. 363–375, 1994.
- [19] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, vol. 59, pp. 731–792, 1997.
- [20] K. Roeder and L. Wasserman, "Practical Bayesian density estimation using mixtures of normals," *J. Amer. Statist. Assoc.*, vol. 92, pp. 894–902, 1997.
- [21] G. Celeux, M. Hurn, and C. P. Robert, "Computational and inferential difficulties with mixture posterior distributions," *J. Amer. Statist. Assoc.*, vol. 95, pp. 957–970, 2000.
- [22] M. Hurn, A. Justel, and C. P. Robert, "Estimating mixtures of regressions," *J. Comput. Graph. Statist.*, vol. 12, pp. 1–25, 2003.
- [23] C. L. Bishop, "Mixture density networks," Dept. Comp. Sci. Applied Math., Aston Univ., Birmingham, U.K., Neural Computing Research Group Rep. NCRG/94/004, 1994.
- [24] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.
- [25] A. Zeevi and R. Meir, "Density estimation through convex combinations of densities; approximation and estimation bounds," *Neural Networks*, vol. 10, pp. 90–109, 1997.
- [26] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 381–396, Mar. 2002.
- [27] J. Hartigan, "A failure of likelihood asymptotics for normal mixtures," in *Proc. Berkeley Conf. Honor of Jerzy Neyman and Jack Kiefer*, vol. II, 1985, pp. 807–810.
- [28] K. Fukumizu, "Likelihood ratio of unidentifiable models and multilayer neural networks," *Ann. Statist.*, vol. 31, pp. 833–851, 2003.
- [29] D. Dacunha-Castelle and E. Gassiat, "The estimation of the order of a mixture model," *Bernoulli*, vol. 3, pp. 279–299, 1997.
- [30] C. E. Priebe, "Adaptive mixtures," *J. Amer. Statist. Assoc.*, vol. 89, pp. 796–806, 1994.

- [31] G. W. Rogers, D. J. Marchette, and C. E. Priebe, "A procedure for model complexity selection in semiparametric mixture model density estimation," Naval Surface Warfare Center, Dahlgren Division, Dahlgren, VA, Tech. Rep., 2002.
- [32] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, vol. 113, pp. 301–413, 1999.
- [33] G. Castellani, "Sélection d'histogrammes à l'aide d'un critère de type Akaike," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 330, pp. 729–732, 2000.
- [34] P. Massart, "Some applications of concentration inequalities to statistics," *Ann. Fac. Sci. Toulouse Math.*, vol. 9, pp. 245–303, 2000.
- [35] J. Rissanen, "On Bayesian analysis of mixtures with an unknown number of components," *Automatica*, vol. 14, pp. 465–471, 1978.
- [36] —, "Stochastic complexity (with discussion)," *J. Roy. Statist. Soc. Ser. B Statist. Methodol.*, vol. 49, pp. 223–239, 1987.
- [37] J. Rissanen, T. Speed, and B. Yu, "Density estimation by stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 38, pp. 315–323, Mar. 1992.
- [38] L. Devroye, *A Course in Density Estimation*. Boston, MA: Birkhäuser, 1987.

Outer Bounds on the Capacity of Gaussian Interference Channels

Gerhard Kramer, *Member, IEEE*

Abstract—Two outer bounds on the capacity region of the two-user Gaussian interference channel (IFC) are derived. The idea of the first bound is to let a genie give each receiver just enough information to decode both messages. This bound unifies and improves the best known outer bounds of Sato and Carleial. Furthermore, the bound extends to discrete memoryless IFCs and is shown to be equivalent to another bound of Carleial. The second bound follows directly from existing results of Costa and Sato and possesses certain optimality properties for weak interference.

Index Terms—Capacity, Gaussian noise, interference.

I. INTRODUCTION

A basic open problem in Shannon theory is the capacity of the interference channel (IFC). A two-user IFC has two transmitters and two receivers, and transmitter t sends its message to receiver t for $t = 1, 2$. The transmissions interfere with each other, and this seems to make the problem hard to solve. The state-of-the-art is today basically the same as at the time of the survey article [1]. The purpose of this correspondence is to unify and improve certain capacity outer bounds.

A memoryless IFC is formally defined by the alphabets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}_1, \mathcal{Y}_2$, and the conditional probability distribution $P(y_1, y_2 | x_1, x_2)$ where $x_t \in \mathcal{X}_t$ and $y_t \in \mathcal{Y}_t$, $t = 1, 2$. The channel is used N times, and is *time invariant* and *memoryless* in the sense that

$$P(y_{1n}, y_{2n} | x_1^n, x_2^n, y_1^{n-1}, y_2^{n-1}) = P_{Y_1 Y_2 | X_1 X_2}(y_{1n}, y_{2n} | x_{1n}, x_{2n}) \quad (1)$$

where x_t^n denotes the sequence $x_{t1}, x_{t2}, \dots, x_{tn}$ and $n = 1, 2, \dots, N$. As in (1), we use the convention of dropping subscripts on the probability distributions when the arguments are the lower case

Manuscript received July 19, 2001; revised January 28, 2003. The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Washington, DC, June 2001.

The author is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: gkr@bell-labs.com).

Communicated by I. E. Telatar, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2004.825249

versions of the random variables. Transmitter t sends a B_t message U_t to receiver t so that the rates are $R_t = B_t/N$ bits per use. The *capacity region* $\mathcal{C}_{\text{IFC}}$ is the closure of the set of rate pairs (R_1, R_2) for which the receivers can decode their messages with arbitrarily small positive error probability.

We will consider primarily the *Gaussian* IFC that has real input and output alphabets, and whose outputs are linear combinations of the transmitted symbols and white Gaussian noise. Furthermore, there are power constraints on the transmissions

$$\sum_{n=1}^N E[X_{tn}^2]/N \leq P_t, \quad t = 1, 2. \quad (2)$$

One can reduce a Gaussian IFC to so-called *standard form* [2], i.e., the received symbols are

$$Y_1 = X_1 + c_{21}X_2 + Z_1 \quad (3)$$

$$Y_2 = c_{12}X_1 + X_2 + Z_2 \quad (4)$$

where c_{21} and c_{12} are real numbers, and Z_1 and Z_2 are independent, zero-mean, unit-variance, Gaussian random variables (see [2, Sec.V-B]).

The capacity of the standard-form Gaussian IFC is currently known only if $c_{21} = c_{12} = 0$ or if $c_{21}^2 \geq 1$ and $c_{12}^2 \geq 1$. The former case is trivial. For the latter case, called *strong interference*, $\mathcal{C}_{\text{IFC}}$ is the set of (R_1, R_2) satisfying

$$0 \leq R_1 \leq \log(P_1 + 1)/2 \quad (5)$$

$$0 \leq R_2 \leq \log(P_2 + 1)/2 \quad (6)$$

$$0 \leq R_1 + R_2 \leq \log(P_1 + c_{21}^2 P_2 + 1)/2 \quad (7)$$

$$0 \leq R_1 + R_2 \leq \log(P_2 + c_{12}^2 P_1 + 1)/2. \quad (8)$$

One can prove this result by the following argument [3]–[5]: for any point (R_1, R_2) inside the capacity region both receivers will, in fact, be able to decode both messages U_1 and U_2 . For example, if receiver 1 can decode U_1 it can determine X_1^N and compute

$$\tilde{Y}_1^N = c_{12}X_1^N + X_2^N + Z_1^N/c_{21}.$$

But $\tilde{Y}_1^N$ is *less noisy* than Y_2^N because $c_{21}^2 \geq 1$. Thus, with high probability, receiver 1 can decode U_2 if receiver 2 can decode U_2 . Repeating this argument for receiver 2 we find that the problem is equivalent to the *compound* IFC, also called the compound multiple-access channel (MAC), that is discussed in [6, Sec.3, p. 14]–[8, p. 288]. The capacity region of this channel is given by (5)–(8). For historical reasons, and because of the remarkable nature of the capacity region, one sometimes distinguishes between *strong* and *very strong* interference. The interference is said to be very strong if the bounds (7) and (8) can be dropped, i.e., if $c_{21}^2 \geq P_1 + 1$ and $c_{12}^2 \geq P_2 + 1$. We refer to [1], [4], [5], [9] for a more detailed discussion of such cases.

This correspondence is organized as follows. In Section II, we review the existing capacity outer bounds for Gaussian IFCs with $0 < c_{21}^2 < 1$ or $0 < c_{12}^2 < 1$. In Section III, we derive an improved bound. Section IV develops another bound that is even better. In Section V, we extend the first bound to discrete memoryless IFCs and recover a bound proposed by Carleial in [10, Theorem 1]. Finally, Section VI suggests possible extensions of our ideas.

II. EXISTING CAPACITY OUTER BOUNDS

Two outer bounds on $\mathcal{C}_{\text{IFC}}$ were described by Sato [11] and Carleial [10]. Sato's bound applies to both discrete memoryless and Gaussian IFCs.