The Annals of Statistics
1995, Vol. 23, No. 4, 1393–1407

A GENERAL CLASSIFICATION RULE FOR PROBABILITY MEASURES[1]

BY SANJEEV R. KULKARNI[2] AND OFER ZEITOUNI[3]

Princeton University and Technion

We consider the composite hypothesis testing problem of classifying an unknown probability distribution based on a sequence of random samples drawn according to this distribution. Specifically, if A is a subset of the space of all probability measures $\mathcal{M}_1(\Sigma)$ over some compact Polish space Σ, we want to decide whether or not the unknown distribution belongs to A or its complement. We propose an algorithm which leads a.s. to a correct decision for any A satisfying certain structural assumptions. A refined decision procedure is also presented which, given a countable collection $A_i \subset \mathcal{M}_1(\Sigma)$, $i = 1, 2, \ldots$, each satisfying the structural assumption, will eventually determine a.s. the membership of the distribution in any finite number of the A_i. Applications to density estimation are discussed.

1. Introduction. In this paper, we consider the composite hypothesis testing problem of classifying an unknown probability distribution into one of a finite or countable number of classes based on random samples drawn from the unknown distribution. This problem arises in a number of applications involving classification and statistical inference. For example, consider the following problems:

1. Given i.i.d. observations $x_1, x_2, \ldots$ from some unknown distribution P, we wish to decide whether the mean of P is in some particular set (e.g., in some interval or whether the mean is rational, etc.).
2. Given i.i.d. observations $x_1, x_2, \ldots$, we wish to decide whether or not the unknown distribution belongs to a particular parametric class (e.g., to determine if it is Gaussian) or to determine to which of a countable hierarchy of classes the unknown distribution belongs (e.g., to determine class membership based on some smoothness parameter of the density function).
3. We wish to decide whether or not observations $x_1, x_2, \ldots$ are coming from a Markov source and, if so, to determine the order of the Markov source.

Received August 1991; revised October 1994.

[1]This work was supported by U.S. Army Research Office Grants DAAL03-86-K-0171 (Center for Intelligent Control Systems) and DAAL03-92-G-0320, by NSF Grants IRI-92-09577 and IRI-94-57645 and by the Office of Naval Research under Air Force Contract F19628-90-C-0002.

[2]This work was partially done while the author was with the Laboratory for Information and Decision Systems, MIT, and with MIT Lincoln Laboratory, Lexington, Massachusetts.

[3]This work was partially done while the author visited the Laboratory for Information and Decision Systems, MIT.

AMS 1991 *subject classifications*. Primary 62F03; secondary 62G10, 62G20.

Key words and phrases. Hypothesis testing, empirical measure, large deviations.

In these examples, our goal is to decide whether an unknown distribution μ belongs to a set of distributions A or its complement A^c, or more generally to decide to which of a countable collection of sets of distributions $A_1, A_2, \ldots$ the unknown μ belongs. After each new observation x_n we will make a decision as to the class membership of the unknown distribution. Our criterion for success is to require that almost surely only a finite number of mistakes are made. There are two aspects to the "almost sure" criterion. First, as expected, we require that with probability 1 (with respect to the observations $x_1, x_2, \ldots$) our decision will be correct from some point on. However, depending on the structure of the A_i, classification may be difficult for certain distributions μ. Hence, given a measure on the set of distributions we allow failure (i.e., do not require a finite number of mistakes) on a set of distributions of measure zero.

Our work is motivated by the previous work of Cover (1973), Koplowitz (1977) and Kulkarni and Zeitouni (1991). In fact, the previous works just mentioned deal with the specific case in which the unknown distribution is to be classified according to its mean based on i.i.d. observations, as in the example problem 1 above. In this case, a subset of $\mathbb{R}$ can be identified with the set of distributions A in the natural way (i.e., all distributions whose mean is in a specified set). Cover (1973) considered the case of distributions on $[0, 1]$ with $A = \mathcal{Q}_{[0,1]}$, the set of rationals in $[0, 1]$ and, more generally, the case of countable A. He provided a test which, for any measure with mean in A or with mean in $A^c \setminus N$, will make (almost surely) only a finite number of mistakes where N is a set of Lebesgue measure 0. For countable A, Cover also considered the countable hypothesis testing problem of deciding exactly the true mean in the case the true mean belongs to A and provided a decision rule satisfying a similar success criterion. Koplowitz (1977) showed some properties of sets A which allow for such decision rules and gave some characterizations of the set N. For example, he showed that if $\bar{A}$ (the closure of A) is countable, then N is empty, while if $\bar{A}$ is uncountable, then N is uncountable. Kulkarni and Zeitouni (1991) extended the results of Cover (1973) by allowing the set A to be uncountable, not necessarily of measure 0, but such that it satisfies a certain structural assumption. Roughly speaking, this structural assumption requires that A be decomposable into a countable union of increasing sets B_m such that a small dilation of B_m increases the Lebesgue measure by only a sufficiently small amount. In a different direction, Dembo and Peres (1994) provide necessary and sufficient conditions for the almost sure discernibility between sets. Their results, when specialized to the setup discussed above, show that the inclusion of the possibility of some errors on the set of irrationals is necessary in order to ensure discernibility.

The decision rules of Cover (1973), Dembo and Peres (1994), Koplowitz (1977) and Kulkarni and Zeitouni (1991) are basically as follows. At time n, the smallest m is selected such that the observations are sufficiently well explained by a hypothesis in B_m. If m is not too large, we decide that the unknown distribution belongs to A; otherwise we decide A^c. For the case of countable hypothesis testing, a similar criterion is used. Thus, the B_m can be thought of as a decomposition of A into hypotheses of increasing complexity

and so the decision rules are reminiscent of Occam's razor or the minimum description length (MDL) principle.

The problem considered in this paper uses a success criterion and decision rules very similar to those in the previous work of Cover (1973), Koplowitz (1977) and Kulkarni and Zeitouni (1991), but allows much more general types of classification of the unknown distribution. Section 2 treats the case of classification in A versus A^c for distributions on an arbitrary compact complete separable metric space (i.e., a compact Polish space) with i.i.d. observations. The case of classification among a countable number of sets $A_1, A_2, \ldots$ from i.i.d. observations is considered in Section 3. Thus, the results of these two sections cover the example problems 1 and 2 mentioned above. Relaxations of the basic assumption concerning the i.i.d. structure of the observations $x_1, \ldots, x_n$ to a Markov situation is possible and rather straightforward. The interested reader is referred to Zeitouni and Kulkarni (1994) for a treatment of example problem 3 on the determination of the order of a Markov chain.

We now give a precise formulation of the problems considered here. Let $x_1, \ldots, x_n$ be i.i.d. samples drawn from some distribution μ. We assume that x_i takes values in some compact Polish space Σ, which for concreteness should be thought of as $[0, 1]^d \subset \mathbb{R}^d$. Let $\mathcal{M}_1(\Sigma)$ denote the space of probability measures on Σ. We put on $\mathcal{M}_1(\Sigma)$ the Prohorov metric, denoted $d(\cdot, \cdot)$, whose topology is equivalent to the weak topology.

We consider here the following problems:

(P1) Based on the sequence of observations $(x_1, \ldots, x_n)$, decide whether $\mu \in A$ or $\mu \in A^c$, where A is some given set satisfying certain structural properties [cf. (A1) below].

(P2) Based on the sequence of observations $(x_1, \ldots, x_n)$, decide whether $\mu \in A_i$, where all $A_i \subset \mathcal{M}_1(\Sigma)$, $i = 1, 2, \ldots$, are sets satisfying structural properties [cf. (A1) below].

Since $\mathcal{M}_1(\Sigma)$ is a Polish space [see Parthasarathy (1967)], there exist on $\mathcal{M}_1(\Sigma)$ many finite Borel measures which we may assume to be normalized to have a total mass 1. Suppose one is given a particular measure, denoted G, on $\mathcal{M}_1(\Sigma)$. In particular, we allow G to charge all open sets in $\mathcal{M}_1(\Sigma)$. The measure G will play the role of the Lebesgue measure in the following structural condition, which is reminiscent of the assumption in Kulkarni and Zeitouni (1991).

(A1) There exists a sequence of open sets $C_m \subset \mathcal{M}_1(\Sigma)$ and closed sets $B_m \subset \mathcal{M}_1(\Sigma)$, and a sequence of positive constants $\varepsilon(m)$ such that the following hold:

(i) $\forall \mu \in A \; \exists \, m_0(\mu) < \infty$ s.t. $\forall \, m > m_0(\mu)$, $\mu \in B_m$;

(ii) $d(B_m, C_m^c) = \sqrt{2\varepsilon(m)} > 0$;

(iii) $G(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} (C_m^{(\sqrt{2\varepsilon(m)})} \setminus A)) = 0$, where $C_m^{(\sqrt{2\varepsilon(m)})} = \{v \in \mathcal{M}_1(\Sigma) \mid d(v, C_m) < \sqrt{2\varepsilon(m)}\}$ is the $\sqrt{2\varepsilon(m)}$ dilation of C_m.

Assumption (A1) is an embellishment of the structural assumption in Kulkarni and Zeitouni (1991), which corresponds to the case where B_m is a monotone sequence and C_m are taken as the $\sqrt{2\varepsilon(m)}$-dilation of B_m. The use of (A1)(i) and (A1)(ii) was proposed to us by A. Dembo and Y. Peres, who obtained also various conditions for full discernibility between hypotheses [cf. Dembo and Peres (1994)]. We note that as in Kulkarni and Zeitouni (1991), the assumption is immediately satisfied for countable sets A by taking as B_m the union of the first m elements of A and noting that, for a finite measure on a metric space, $G(B(x, \delta) \setminus \{x\}) \to_{\delta \to 0} 0$, where $B(x, \delta)$ denotes the open ball of radius δ around x. More generally, (A1) is satisfied for any closed set by taking $B_m = A$ and using for C_m a sequence of open sets which include A whose measure converges to the outer measure of A. Since C_m is open and Σ is compact, it follows that $d(A, C_m^c) > 0$, and (A1) is satisfied. By the same considerations, it also follows that (A1) is satisfied for any countable union of closed sets. Also, note that whenever both $A = \bigcup_{i=1}^{\infty} A_i$ and $A^c = \bigcup_{i=1}^{\infty} D_i$, with A_i and D_i closed, then, choosing $B_m = \bigcup_{i=1}^{m} A_i$ and $C_m = \bigcap_{i=1}^{m} D_i^c$, one sees that (A1) holds [with actually an empty intersection in (A1)(iii) for appropriate $\varepsilon(m)$]. In this situation, the results of this paper correspond to the sufficient part of Dembo and Peres (1994).

2. Classification in A versus A^c. The definition of success of the decision rule will be similar to the one used in Kulkarni and Zeitouni (1991). Namely, a test which makes at each instant n a decision whether $\mu \in A$ or $\mu \in A^c$ based on $x_1, \ldots, x_n$ will be called *successful* if the following hold:

(S1) $\forall \mu \in A$, a.s. ω, $\exists\, T(\omega)$ s.t. $\forall\ n > T(\omega)$, the decision is A;
(S2) $\exists\, N \subset \mathcal{M}_1(\Sigma)$ s.t.
  (S2.1) $G(N) = 0$,
  (S2.2) $\forall \mu \in A^c \setminus N$, a.s. ω, $\exists\, T(\omega)$ s.t. $\forall\, n > T(\omega)$, the decision is A^c.

Note that the outcome is unspecified on N. Note also that the definition is asymmetric in the roles played by A and A^c in the sense that errors in A are not allowed at all.

Let $\mu_n = (1/n)\sum_{i=1}^{n} \delta_{x_i}$. We recall that μ_n satisfies a large-deviation principle, that is,

$$
\begin{aligned}
(1) \qquad -\inf_{\theta \in A^0} H(\theta \mid \mu) &\le \liminf_{n \to \infty} \frac{1}{n} \log P(\mu_n \in A) \\
&\le \limsup_{n \to \infty} \frac{1}{n} \log P(\mu_n \in A) \le -\inf_{\theta \in \bar{A}} H(\theta \mid \mu),
\end{aligned}
$$

where $\bar{A}$ (A^0) denotes the closure (interior) of a set $A \subset \mathcal{M}_1(\Sigma)$ in the weak topology, respectively, and

$$
(2) \qquad H(\theta \mid \mu) = \begin{cases} \int_{\Sigma} d\theta \log \dfrac{d\theta}{d\mu}, & \text{if } \theta \ll \mu, \\ \infty, & \text{otherwise.} \end{cases}
$$

[See Deuschel and Stroock (1989) or Dembo and Zeitouni (1993) for an introduction to the theory of large deviations.]

Our decision rule is very similar to that in Kulkarni and Zeitouni (1991). Specifically, we parse the input sequence $x_1, x_2, \ldots$ to form the subsequences

$$X^m \triangleq (x_{\beta(m-1)+1}, \ldots, x_{\beta(m)}), \tag{3}$$

where the choice of the $\beta(m)$ will be given below. The length of the sequence X^m will be denoted by $\alpha(m)$, so that

$$\beta(m) = \sum_{i=1}^{m} \alpha(i), \qquad \beta(0) = 0. \tag{4}$$

We will specify the $\beta(m)$ by appropriately selecting the lengths $\alpha(m)$ of the subsequences.

At the end of each subsequence X^m, we form the empirical measure μ_{X^m} based on the data in the subsequence X^m, namely,

$$\mu_{X^m} = \frac{1}{\alpha(m)} \sum_{i=\beta(m-1)+1}^{\beta(m)} \delta_{x_i}. \tag{5}$$

Then we make a decision of whether $\mu \in A$ or $\mu \in A^c$ according to whether $\mu_{X^m} \in C_m$ or not. Between parsings, we do not change the decision.

Recall that, from the structural assumption (A1), C_m^c is $\sqrt{2\varepsilon(m)}$-separated from B^m. We choose $\alpha(m)$ sufficiently large such that if the true measure μ is in B_m, then we will have enough data in forming the empirical measure μ_{X^m} to make the probability of an incorrect decision (deciding A^c because $\mu_{X^m} \in C_m^c$) less than $1/m^2$. If $\alpha(m)$ can be chosen in this manner, then, for any $\mu \in A$, once $m > m_0(\mu)$ our probability of error at the end of the parsing interval m is less than $1/m^2$ so that by the Borel–Cantelli lemma we make only finitely many errors.

To show that $\alpha(m)$ can be chosen to satisfy the necessary properties, we will need a strengthened version of the upper bound in Sanov's theorem (1). To do that, we use the notion of covering numbers.

DEFINITION. Let $\varepsilon > 0$ be given. The *covering number* of $\mathcal{M}_1(\Sigma)$, denoted $N(\varepsilon, \mathcal{M}_1(\Sigma))$, is defined by

$$N(\varepsilon, \mathcal{M}_1(\Sigma)) \triangleq \inf\left\{ n \mid \exists\, y_1, \ldots, y_n \in \mathcal{M}_1(\Sigma) \text{ s.t. } \mathcal{M}_1(\Sigma) \subset \bigcup_{i=1}^{n} B(y_i, \varepsilon) \right\}, \tag{6}$$

where $B(y, \varepsilon)$ denotes a ball of radius ε (in the Prohorov metric) around y.

Similarly, for any given $\varepsilon > 0$, denote by $N^{\Sigma}(\varepsilon)$ the covering number of Σ, that is,

$$N^{\Sigma}(\varepsilon) \triangleq \inf\left\{ n \mid \exists\, \tilde{y}_1, \ldots, \tilde{y}_n \in \Sigma \text{ s.t. } \Sigma \subset \bigcup_{i=1}^{n} B(\tilde{y}_i, \varepsilon) \right\}. \tag{7}$$

where $B(\tilde{y}_i, \varepsilon)$ are taken in the metric corresponding to Σ.

We now make the following claim.

LEMMA 1.

$$N(\varepsilon, \mathcal{M}_1(\Sigma)) \le 2\left(\frac{e}{\varepsilon}\right)^{(N^\Sigma(\varepsilon))} \triangleq \bar{N}(\varepsilon, \mathcal{M}_1(\Sigma)). \tag{8}$$

PROOF. In order to prove the lemma, we will explicitly construct an ε-cover of $\mathcal{M}_1(\Sigma)$ with $\bar{N}(\varepsilon, \mathcal{M}_1(\Sigma))$ elements.

Let $\tilde{y}_1, \ldots, \tilde{y}_{N^\Sigma(\varepsilon)}$ be the centers of a set of ε balls in Σ which create the cover $N^\Sigma(\varepsilon)$ in (7). Let $\delta_i \triangleq \delta_{\tilde{y}_i}$, that is, the distribution concentrated at $\tilde{y}_i$, and let

$$\mu_i^j \triangleq j \cdot \left(\frac{\varepsilon}{N^\Sigma(\varepsilon)}\right) \cdot \delta_i, \qquad j = 0, 1, \ldots, \frac{N^\Sigma(\varepsilon)}{\varepsilon}.$$

Define

$$Y \triangleq \left\{ y \in \mathcal{M}_1(\Sigma) : \exists (i_1, j_1), \ldots, (i_k, j_k) \text{ s.t. } y = \sum_{\alpha=1}^{k} \mu_{i_\alpha}^{j_\alpha} \right\}.$$

Note that Y is a finite set, for it includes at most $((N^\Sigma(\varepsilon)/\varepsilon) + 1)^{N^\Sigma(\varepsilon)}$ members. Also, note that Y is an ε-cover of $\mathcal{M}_1(\Sigma)$, that is, for any $\mu \in \mathcal{M}_1(\Sigma)$ there exists a $y \in Y$ such that, for any open set $C \subset \Sigma$, $\mu(C) \le y(C^\varepsilon) + \varepsilon$. To see that, choose as y the following approximation to μ. Let $i_\alpha = \alpha$, $\alpha = 1, \ldots, N^\Sigma(\varepsilon)$, and choose

$$j_\alpha = \left\lfloor \mu\left(B(\tilde{y}_\alpha, \varepsilon) \setminus \left(\bigcup_{k=1}^{\alpha-1} B(\tilde{y}_k, \varepsilon) \right) \right) \right\rfloor \frac{N^\Sigma(\varepsilon)}{\varepsilon},$$

where by $\lfloor \times \rfloor$ we mean the closest approximation to $\times$ on the $N^\Sigma(\varepsilon)/\varepsilon$ j-net from below. Finally, let

$$j_{N^\Sigma(\varepsilon)} \triangleq \frac{N^\Sigma}{\varepsilon} - \sum_{\alpha=1}^{N^\Sigma(\varepsilon)-1} j_\alpha.$$

Now take $y = \sum_{\alpha=1}^{N^\Sigma(\varepsilon)} \mu_{i_\alpha}^{j_\alpha}$. Due to our choice of $j_{N^\Sigma(\varepsilon)}$, it follows that y is a probability measure based on a finite number of atoms. Furthermore, since for every measurable set C one has

$$\mu\left(C \cap \left(B(\tilde{y}_\alpha, \varepsilon) \setminus \left(\bigcup_{k=1}^{\alpha-1} B(\tilde{y}_k, \varepsilon) \right) \right) \right) \le \mu_{i_\alpha}^{j_\alpha} + \frac{\varepsilon j_\alpha}{N^\Sigma(\varepsilon)},$$

it follows by construction that for every open (actually, for every measurable) set C one has that $\mu(C) \le y(C^\varepsilon) + \varepsilon$ and hence, by definition, $d(y, \mu) < \varepsilon$ [recall that the metric on $\mathcal{M}_1(\Sigma)$ is Prohorov]. We need therefore only to estimate the cardinality of the set Y, denoted $|Y|$. Note that $|Y|$ is just the

number of vectors $(j_1, \ldots, j_{N^\Sigma(\varepsilon)})$ such that $\sum_{i=1}^{N^\Sigma(\varepsilon)} j_i = 1$ and $j_i \in \{0, \varepsilon/N(\varepsilon), 2\varepsilon/N(\varepsilon), \ldots, 1\}$. Hence,

$$
\begin{aligned}
|Y| &\le \left(\frac{N^\Sigma(\varepsilon)}{\varepsilon} + 1\right)^{N^\Sigma(\varepsilon)} \int_0^1 \cdots \int_0^{x_3} \int_0^{x_2} dx_1 \cdots dx_{N^\Sigma(\varepsilon)} \\
&= \left(\frac{N^\Sigma(\varepsilon)}{\varepsilon} + 1\right)^{N^\Sigma(\varepsilon)} \cdot \frac{1}{N^\Sigma(\varepsilon)!}
\end{aligned} \tag{9}
$$

However, by Stirling's formula,

$$
\log(N^\Sigma(\varepsilon)!) \ge N^\Sigma(\varepsilon)\log N^\Sigma(\varepsilon) - N^\Sigma(\varepsilon). \tag{10}
$$

Substituting (10) into (9), one has

$$
|Y| \le \left(\frac{N^\Sigma(\varepsilon)}{\varepsilon} + 1\right)^{N^\Sigma(\varepsilon)} \exp\left[N^\Sigma(\varepsilon)\right] \cdot \frac{1}{\left(N^\Sigma(\varepsilon)\right)^{N^\Sigma(\varepsilon)}}, \tag{11}
$$

which implies that

$$
\begin{aligned}
N(\varepsilon, \mathcal{M}_1(\Sigma)) &\le \left(\frac{1}{\varepsilon}\left(1 + \frac{\varepsilon}{N^\Sigma(\varepsilon)}\right)\right)^{N^\Sigma(\varepsilon)} \exp\left[N^\Sigma(\varepsilon)\right] \\
&= \left(\frac{e}{\varepsilon}\right)^{N^\Sigma(\varepsilon)} \left(1 + \frac{\varepsilon}{N^\Sigma(\varepsilon)}\right)^{N^\Sigma(\varepsilon)} \le 2\left(\frac{e}{\varepsilon}\right)^{N^\Sigma(\varepsilon)} \\
&= \overline{N}(\varepsilon, \mathcal{M}_1(\Sigma)).
\end{aligned}
$$

□

For completeness, we show in the Appendix a complementary lower bound on the covering number which exhibits a behavior similar to $\overline{N}$. Thus, the upper bound $\overline{N}$ cannot be much improved.

The existence of the bound $\overline{N}$ permits us to mimic the computation in Kulkarni and Zeitouni (1991) for the case in hand. Indeed, a crucial step needed is bounding the probability of complements of balls, for all n, uniformly over all measures, as follows.

THEOREM 1.

$$
P\left(\mu_n \in B(\mu, \delta)^c\right) \le \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \exp\left[-n\left(\frac{\delta}{4}\right)^2\right].
$$

PROOF. The proof follows the standard Chebyshev bound technique, without taking n limits as in the large-deviation framework. Indeed,

$$
P\left(\mu_n \in B(\mu, \delta)^c\right) \le \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \cdot \sup_{y \in \mathcal{M}_1(\Sigma),\, d(y,\mu) \ge 3\delta/4} P\left(\mu_n \in B\left(y, \frac{\delta}{4}\right)\right).
$$

Therefore, by the Chebyshev bound, letting P_n denote the law of the random variable μ_n and letting $C_b(\Sigma)$ denote the space of continuous (and hence, by compactness, bounded) functions on Σ, it follows that, for any $\theta \in C_b(\Sigma)$,

$$
\begin{aligned}
&P\big(\mu_n \in B(\mu, \delta)^c\big) \\
&\quad \le \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \\
&\qquad \times \sup_{y \in \mathcal{M}_1(\Sigma),\, d(y,\mu) \ge 3\delta/4} \int_{B(y,\delta/4)} \exp(n\langle\theta, \nu\rangle)\exp(-n\langle\theta, \nu\rangle)\, dP_n(\nu) \\
&\quad \le \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \\
&\qquad \times \sup_{y:\, d(y,\mu) \ge 3\delta/4} \exp\left(-n \sup_{\theta \in C_b(\Sigma)} \inf_{\nu \in B(y,\delta/4)} \left(\langle\theta, \nu\rangle \right.\right. \\
&\qquad\qquad \left.\left. - \frac{1}{n}\log E_{P_n}(\exp[\, n\langle\theta, \nu\rangle])\right)\right) \\
(12) \quad &\quad = \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \\
&\qquad \times \exp\left(-n \inf_{\nu \in B(y,\delta/4),\, d(y,\mu) \ge 3\delta/4} \sup_{\theta \in C_b(\Sigma)} \left(\langle\theta, \nu\rangle \right.\right. \\
&\qquad\qquad \left.\left. - \frac{1}{n}\log E_{P_n}(\exp[\, n\langle\theta, \nu\rangle])\right)\right) \\
&\quad = \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right)\exp\left(-n \inf_{\nu \in B(y,\delta/4),\, d(y,\mu) \ge 3\delta/4} H(\nu \mid \mu)\right) \\
&\quad \le \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right)\exp\left(-n \inf_{\nu \in B(\mu,\delta/2)^c} H(\nu \mid \mu)\right) \\
&\quad \le \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right)\exp\left[-n\left(\frac{\delta}{4}\right)^2\right],
\end{aligned}
$$

where $\langle\theta, \nu\rangle = \int \theta(x)\nu(dx)$, the first equality in (12) follows from the min–max theorem for convex compact sets [cf. Sion (1958), Theorem 4.2], the second equality follows by Lemma 3.2.13 of Deuschel and Stroock (1989), and the last inequality from the fact that [Deuschel and Stroock (1989), Exercise 3.2.24], for any $\eta \in B(\mu, \delta/2)^c$,

$$
\frac{\delta}{2} \le d(\eta, \mu) \le \|\eta - \mu\|_{\mathrm{var}} \le 2H^{1/2}(\eta \mid \mu). \qquad \square
$$

COROLLARY 1. *Let $B_m \subset \mathcal{M}_1(\Sigma)$ be a measurable set such that $\mu \in B_m$. Let B_m^δ denote an open set such that $d(B_m, (B_m^\delta)^c) \geq \delta$. Then*

$$P\left(\mu_m \in \left(B_m^\delta\right)^c\right) \leq \overline{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right)\exp\left[-n\left(\frac{\delta}{4}\right)^2\right]. \tag{13}$$

We return now to the proposed classification algorithm. Motivated by Corollary 1, define

$$\alpha(m) = \frac{8}{\varepsilon(m)}\left[2\log m + \log 2 + N^\Sigma\left(\sqrt{\frac{\varepsilon(m)}{8}}\right)\left(1 - \log\sqrt{\frac{\varepsilon(m)}{8}}\right)\right] \tag{14}$$

and let $\beta(m)$ be as defined previously by (4).

Note that with this choice of $\alpha(m)$, using Corollary 1 with $\delta = \sqrt{2\varepsilon(m)}$, and the expression for $\overline{N}(\delta/4, \mathcal{M}_1(\Sigma))$ from Lemma 1, we have that, for all $\mu \in A$ and $m > m_0(\mu)$,

$$P\left(\mu_{\alpha(m)} \in C_m^c\right) \leq \frac{1}{m^2}, \tag{15}$$

as we wanted.

For convenience we summarize the decision rule again here.

DECISION RULE. For any input sequence $x_1, x_2, \ldots$, form the subsequences

$$X^m \hat{=} \left(x_{\beta(m-1)+1}, \ldots, x_{\beta(m)}\right).$$

Let μ_{X^m} denote the empirical measure of the sequence X^m. At the end of each parsing, decide $\mu \in A$ if $\mu_{X^m} \in C_m$, and decide $\mu \in A^c$ otherwise. Between parsings, do not change the decision.

We now make the following claim.

THEOREM 2. *The decision rule defined by the parsing $\beta(m)$ as above is successful.*

PROOF. The proof is essentially identical to the proof of Theorem 1 in Kulkarni and Zeitouni (1991).

(a) If $\mu \in A$, then by assumption (A1)(i) there exists $m_0(\mu)$ such that $\mu \in B_m$ for all $m > m_0(\mu)$. Note that the event of making an error infinitely often is equivalent to the event of making an error *at the parsing intervals* infinitely often. However, by our choice of $\alpha(m)$,

$$\sum_{m=1}^{\infty} \text{Prob}\{\text{error after } m\text{th parsing}\} \leq m_0(\mu) + \sum_{m=m_0(\mu)+1}^{\infty} \frac{1}{m^2} < \infty.$$

Therefore, using the Borel–Cantelli lemma, we have that our decision rule satisfies part (S1) of the definition of a successful decision rule.

(b) Let

$$N = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} C_m^{(\sqrt{2\varepsilon(m)})} \setminus A. \tag{16}$$

By assumption (A1)(iii), $G(N) = 0$. Now, if $\mu \in A^c \setminus N$, we may repeat the arguments of part (a) in the following way: For an $m_0(\mu)$ large enough, $\mu \in (C_m^{(\sqrt{2\varepsilon(m)})})^c$ for all $m > m_0(\mu)$. Therefore, we have $d(\mu, C_m) \geq \sqrt{2\varepsilon(m)}$ for all $m > m_0(\mu)$. Then, using Corollary 1 with $\delta = \sqrt{2\varepsilon(m)}$, the expression for $\bar{N}(\delta/4, \mathcal{M}_1(\Sigma))$ from Lemma 1 and the choice of $\alpha(m)$, we have that, for $m > m_0(\mu)$,

$$\text{Prob}\{\text{error after } m\text{th parsing}\} = P(\mu_{X^m} \in C_m) \leq \frac{1}{m^2}. \tag{17}$$

Hence, as in part (a), the result follows by a simple application of the Borel–Cantelli lemma. □

3. Classification among a countable number of sets. In this section, we refine the decision rule to allow for classification among a countable number of sets. Specifically, if $A_1, A_2, \ldots$ are a countable number of subsets of $\mathcal{M}^1(\Sigma)$, we are interested in deciding to which of the A_i the unknown measure μ belongs. The only assumption we make on the A_i is that each A_i satisfies the structural assumption (A1). The A_i are not required to be either disjoint or nested, although these special cases are most commonly of interest in applications. In general, after a finite number of observations one cannot expect to determine the membership status of μ in all of the A_i. However, we will show that for all μ except in a set of G-measure zero in $\mathcal{M}^1(\Sigma)$ there is a decision procedure that a.s. will eventually determine the membership of μ in any finite subset of the A_i. In the special cases of disjoint or nested A_i, the membership status of μ in any of the countable A_i is completely determined by membership in some finite subset. Hence, in these cases, except for μ in a set of G-measure zero the membership of μ in all the A_i will a.s. be eventually determined.

We modify our previous decision rule as follows. The observations $x_1, x_2, \ldots$ will still be parsed into increasingly larger blocks in a manner to be defined below. However, now, at the end of the mth block, we will make a decision as to the membership of μ in the first m of the A_i. The decisions of whether μ belongs to $A_1, \ldots, A_m$ are made separately for each A_i using a procedure similar to that of the previous section.

Specifically, for each A_i, let $B_{i,m}$ be a sequence of closed sets, let $C_{i,m}$ be a sequence of open sets and let $\varepsilon_i(m) \to_{m\to\infty} 0$ be a positive sequence satisfying

the requirements of the structural assumption (A1). From the same considerations that led to (15), for

$$\alpha_i(m) = \frac{8}{\varepsilon_i(m)}\Big[2\log m + \log 2 + N^{\Sigma}\Big(\sqrt{\varepsilon_i(m)/8}\Big)\Big(1 - \log\sqrt{\varepsilon_i(m)/8}\Big)\Big], \tag{18}$$

we have, for $\mu \in A_i$,

$$P_\mu\big(\mu_{\alpha_i(m)} \in C_{i,m}^c\big) \le \frac{1}{m^2}. \tag{19}$$

As before, the observation sequence $x_1, x_2, \ldots$ will be parsed into nonoverlapping blocks

$$X^m = (x_{\beta(m-1)+1}, \ldots, x_{\beta(m)}), \tag{20}$$

where the $\beta(m)$ are defined below. At the end of the mth block, a decision will be made about the membership of μ in $A_1, \ldots, A_m$. This decision will be made separately for each $i = 1, \ldots, m$ using the observation sequence X^m exactly as before: that is, at the end of the parsing sequence X^m, for $i = 1, \ldots, m$, decide that $\mu \in A_i$ according to whether or not $\mu_{X^m} \in C_{i,m}$, and do not change the decision except at the end of a parsing sequence. We define the parsing sequence $\beta(m)$ by $\beta(0) = 0$ and $\beta(m) - \beta(m-1) = \max_{1 \le i \le m} a_i(m)$ or, equivalently,

$$\beta(m) = \sum_{k=1}^{m} \max_{1 \le i \le k} \alpha_i(k), \qquad \beta(0) = 0. \tag{21}$$

For this decision rule we have the following theorem.

THEOREM 3. *Let $A_i \subset \mathscr{M}^1(\Sigma)$, for $i = 1, 2, \ldots$, satisfy the structural assumption (A1). Then there is a set $N \subset M_1(\Sigma)$ of G-measure zero such that, for every $\mu \in \mathscr{M}^1(\Sigma) \setminus N$ and every $k < \infty$, the decision rule will make (a.s.) only a finite number of mistakes in deciding the membership of μ in $A_1, \ldots, A_k$: that is, given any $\mu \in \mathscr{M}^1(\Sigma) \setminus N$, for a.e. ω there exists $m(\omega) = m(\omega, \mu, k)$ such that for all $m > m(\omega)$ the algorithm makes a correct decision as to whether $\mu \in A_i$ or $\mu \in A_i^c$, for $i = 1, \ldots, k$.*

PROOF. Let

$$N_i = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} C_{i,m}^{(\sqrt{2\varepsilon(m)})} \setminus A_i, \tag{22}$$

and let

$$N = \bigcup_{i=1}^{\infty} N_i. \tag{23}$$

Then from assumption (A1) it follows that the G-measure of each N_i is zero, and so the G-measure of N is also zero.

Now, let $\mu \in \mathcal{M}^1(\Sigma) \setminus N$, and let $k < \infty$. For each $i = 1, \ldots, k$, there exists $m_i(\mu) < \infty$ such that if $\mu \in A_i$, then $\mu \in B_{i,m}$ for all $m > m_i(\mu)$, while if $\mu \in A_i^c$, then $\mu \in (C_{i,m}^{(\sqrt{2\varepsilon(m)})})^c$ for all $m > m_i(\mu)$ (since $\mu \notin N_i$). Recall that, at the end of the parsing sequence X^m, the algorithm decides $\mu \in A_i$ iff $\mu_{X^m} \in C_{i,m}$, so that if $\mu \in A_i$, then an error is made about membership in A_i iff $\mu_{X^m} \notin C_{i,m}$ while if $\mu \notin A_i$, an error is made iff $\mu_{X^m} \in C_{i,m}$. If $\mu \in A_i$, then, using Corollary 1 and the fact that $d(B_{i,m}, C_{i,m}^c) \geq \sqrt{2\varepsilon_i(m)}$, we have that the probability of making an incorrect decision is less than $1/m^2$ for $m > m_i(\mu)$. On the other hand, if $\mu \in A_i^c$, then since

$$d\left(C_{i,m}, \left(C_{i,m}^{(\sqrt{2\varepsilon(m)})}\right)^c\right) \geq \sqrt{2\varepsilon_i(m)}$$

we also have probability of error less than $1/m^2$ for $m > m_i(\mu)$ [again using Corollary 1 and the expression for $\alpha(m)$]. Hence, for $m > m_0(\mu) = \max(m_1(\mu), \ldots, m_k(\mu))$, the probability of making an error about the membership of μ in *any* of $A_1, \ldots, A_k$ is less than k/m^2. Then

$$\sum_{m=1}^{\infty} \text{Prob}\{\text{error in any } A_i \text{ on } m\text{th parsing}\} \leq m_0 + k \sum_{m=m_0+1}^{\infty} \frac{1}{m^2} < \infty,$$

so that the theorem follows by the Borel–Cantelli lemma. □

Note that if one also wants to make a correct decision after some finite time whether or not μ is in *any* of the A_i, for $i = 1, 2, \ldots$, then the decision procedure can be easily modified to handle this. Specifically, it is easy to show that sets satisfying the structural assumption are closed under countable union. Hence, one could include in the hypothesis testing the set $A_0 = \bigcup_{i=1}^{\infty} A_i$, so that after some finite time a correct decision would be made about the membership of $\mu \in A_0$.

Also, it is worthwhile to note that if the A_i have more structure then some improvements can be made. For example, if the membership status of μ in A_i, for $i = 1, 2, \ldots$, is determined by its membership status in some finite number of the A_i, then a correct decision regarding the membership of μ in all of the A_i can be guaranteed (a.s.) after some finite time (depending on μ). This is the case for disjoint or nested A_i, which may be of particular interest in some applications. For these cases, by letting $A_0 = \bigcup_{i=1}^{\infty} A_i$ and running the decision rule on $A_0, A_1, A_2, \ldots$ as mentioned above, we have the following corollary of Theorem 3.

COROLLARY 2. *Let $A_i \subset \mathcal{M}^1(\Sigma)$, for $i = 1, 2, \ldots$, satisfy the structural assumption (A1) and suppose the A_i are either disjoint or nested. Then there is a set $N \subset M_1(\Sigma)$ of G-measure zero such that, for every $\mu \in \mathcal{M}^1(\Sigma) \setminus N$, the decision rule will make (a.s.) only a finite number of mistakes in deciding the*

membership of μ in all of the A_i: that is, given any $\mu \in \mathscr{M}^1(\Sigma) \setminus N$, for a.e. μ there exists $m(\omega) = m(\omega, \mu)$ such that for all $m > m(\omega)$ the algorithm makes a correct decision as to whether $\mu \in A_i$, for all $i = 1, 2, \ldots$.

It is worthwhile to note that the results of this section may be used also in the case that Σ is locally compact but not compact. In that case, one may first intersect the A_i with compact sets K_m which sequentially approximate Σ and then use $m(n) \to \infty$. We do not consider this issue here.

We conclude this section with an example taken from the problem of density estimation. Let $\Sigma = [0, 1]$ and assume that $x_1, \ldots, x_n$ are i.i.d. and drawn from a distribution with law μ_θ, $\theta \in \Theta$. When some structure is given on the set $\mathscr{F} = \bigcup_{\theta \in \Theta} \mu_\theta$, there exists a large body of literature which enables one to obtain estimates of the error after n observations [e.g., see Ibragimov and Has'minskii (1981)]. All these results assume an a priori structure, for example, a bound on the L^2-norm of the density $f_\theta = d\mu_\theta/dx$. If such information is not given a priori, it may be helpful to design a test to check for this information and thus to be able to estimate eventually whether the distribution belongs to a nice set and if so, to apply the error estimates alluded to above. The application of such an idea to density estimation was suggested by Cover (1972).

As a specific example, let

$$A_i = \left\{ \mu \in \mathscr{M}_1(\Sigma) : \int_0^1 \left(\frac{d\mu(x)}{dx} \right)^2 \leq i \right\}.$$

Note that the sets A_i are closed w.r.t. the Prohorov metric and therefore they satisfy the structural assumption (A1). Moreover, they are nested and thus Corollary 2 may be applied to yield a decision rule which will asymptotically decide correctly on the appropriate class of densities.

A somewhat different application to density estimation arises when the A_i consist of single points (i.e., each A_i contains a single probability measure). The special case in which A_i consists of the ith computable density is related to a model considered by Barron (1985) and Barron and Cover (1991). For an estimation procedure based on the minimum description length (MDL) principle, they showed strong consistency results when the true density is a computable one. Since there are a countable number of computable densities and the structural assumption (A1) is satisfied for any singleton, a strong consistency result for computable densities follows immediately from our results.

APPENDIX

For completeness, here we prove a lower bound for the covering number of $\mathscr{M}_1(\Sigma)$ with respect to the Prohorov metric. This lower bound exhibits a behavior similar to the upper bound of (8), so that these bounds cannot be much improved. In the proof below, $M(\varepsilon, Y, \eta)$ denotes the ε-capacity (or

packing number) of the space Y with respect to the metric η, that is, $M(\varepsilon, Y, \eta)$ represents the maximum number of nonoverlapping balls of *diameter* ε with respect to the metric η that can be packed in Y. The well-known relationship

$$N(2\varepsilon, Y, \eta) \leq M(2\varepsilon, Y, \eta) \leq N(\varepsilon, Y, \eta)$$

between covering numbers and packing numbers is easy to show and is used in the proof below. Note that, for a Polish space Σ with metric η, we use the notation $N(\varepsilon, \Sigma, \eta) = N^{\Sigma}(\varepsilon)$ and $N(\varepsilon, \mathcal{M}^1(\Sigma), d) = N(\varepsilon, \mathcal{M}^1(\Sigma))$.

LEMMA A1. *Let Σ be compact Polish space with metric η, and let $\mathcal{M}^1(\Sigma)$ denote the set of probability measures on Σ with the Prohorov metric d. Then*

$$N(\varepsilon, \mathcal{M}^1(\Sigma)) \geq 8\varepsilon\sqrt{N^{\Sigma}(2\varepsilon)}\left(\frac{1}{8\varepsilon}\right)^{N^{\Sigma}(2\varepsilon)}.$$

PROOF. First, we can find $N = N^{\Sigma}(\varepsilon)$ points $x_1, \ldots, x_n$ which are pairwise greater than or equal to ε apart. Each measure supported on these N points corresponds to a point in $\mathbb{R}^N$ in the natural way. Then, the set of all probability measures supported on $x_1, \ldots, x_N$ corresponds to the simplex S^N in $\mathbb{R}^N$.

Now, let p and q be points on the simplex S^N and suppose that $d_{\ell^1}(p, q) \geq 2\varepsilon$, where $d_{\ell^1} = \sum_{i=1}^{N} |p_i - q_i|$. Then on some subset $G \subset \{1, \ldots, N\}$ of coordinates either $\sum_{i \in G} p_i \leq \sum_{i \in G} q_i + \varepsilon$ or $\sum_{i \in G} q_i \leq \sum_{i \in G} p_i + \varepsilon$. Then, considered as probability measures on Σ, $d(p, q) \geq \varepsilon$ since there is a closed set $F \subset \Sigma$, namely, $F = \{x_i | i \in G\}$, for which either $p(F) \geq q(F^{\varepsilon}) + \varepsilon$ or $q(F) \geq p(F^{\varepsilon}) + \varepsilon$. Hence,

$$N(\varepsilon/2, \mathcal{M}^1(\Sigma), d) \geq M(\varepsilon, \mathcal{M}^1(\Sigma), d) \geq M(2\varepsilon, S^N, d_{\ell^1}) \geq N(2\varepsilon, S^N, d_{\ell^1}).$$

Finally, to get a lower bound on $N(2\varepsilon, S^N, d_{\ell^1})$, we note that the $(N-1)$-dimensional surface measure of the simplex S^N is $\sqrt{N}/(N-1)!$ (simply, differentiate the N-dimensional volume of the interior of an x-scaled simplex with respect to x, taking the angles into account). On the other hand, note that the $(N-1)$-dimensional volume of the intersection of S^N with an N-dimensional ℓ^1-ball of radius 2ε is not larger than the volume of an $(N-1)$-dimensional ℓ^1-ball of radius 2ε, which equals $(4\varepsilon)^{N-1}/(N-1)!$. Thus, $N(2\varepsilon, S^N, d_{\ell^1}) \geq (1/4\varepsilon)^{N-1}\sqrt{N}$. Thus,

$$N\left(\frac{\varepsilon}{2}, \mathcal{M}^1(\Sigma)\right) \geq \left(\frac{1}{4\varepsilon}\right)^{N^{\Sigma}(\varepsilon)} 4\varepsilon\sqrt{N^{\Sigma}(\varepsilon)},$$

or equivalently,

$$N(\varepsilon, \mathcal{M}^1(\Sigma)) \geq 8\varepsilon\sqrt{N^{\Sigma}(2\varepsilon)}\left(\frac{1}{8\varepsilon}\right)^{N^{\Sigma}(2\varepsilon)}.$$

REFERENCES

BARRON, A. R. (1985). Logically smooth density estimation. Ph.D. dissertation, Dept. Electrical Engineering, Stanford Univ.

BARRON, A. R. and COVER, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054.

COVER, T. M. (1972). A hierarchy of probability density function estimates. In *Frontiers of Pattern Recognition* 83–98. Academic Press, New York.

COVER, T. M. (1973). On determining the irrationality of the mean of a random variable. *Ann. Statist.* **1** 862–871.

DEMBO, A. and PERES, Y. (1994). A topological criterion for hypothesis testing. *Ann. Statist.* **22** 106–117.

DEMBO, A. and ZEITOUNI, O. (1993). *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston.

DEUSCHEL, J. D. and STROOCK, D. W. (1989). *Large Deviations*. Academic Press, New York.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation*. Springer, New York.

KOPLOWITZ, J. (1977). Abstracts of papers. In *IEEE International Symposium on Information Theory* 64. IEEE, New York.

KULKARNI, S. R. and ZEITOUNI, O. (1991). Can one decide the type of the mean from the empirical measure? *Statist. Probab. Lett.* **12** 323–327.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.

SION, M. (1958). On general minimax theorems. *Pacific J. Math.* **8** 171–175.

ZEITOUNI, O. and KULKARNI, S. R. (1994). A general classification rule for probability measures. Report LIDS-P-2272, Laboratory for Information and Decision Systems, MIT.

SANJEEV R. KULKARNI
DEPARTMENT OF ELECTRICAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544

OFER ZEITOUNI
DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION
HAIFA 32000
ISRAEL