ALMOST SURE CLASSIFICATION OF DENSITIES

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2A7

Gábor Lugosi
Department of Economics
Pompeu Fabra University
Ramon Trias Fargas, 25-27
08005 Barcelona, Spain

Abstract. Let a class $\mathcal{F}$ of densities be given. We draw an i.i.d. sample from a density f which may or may not be in $\mathcal{F}$. After every n, one must make a guess whether $f \in \mathcal{F}$ or not. A class is almost surely discernible if there exists such a sequence of classification rules such that for any f, we make finitely many errors almost surely. In this paper several results are given that allow one to decide whether a class is almost surely discernible. For example, continuity and square integrability are not discernible, but unimodality, log-concavity, and boundedness by a given constant are.

Keywords and phrases. Density estimation, kernel estimate, convergence, discernibility, hypothesis testing, asymptotic optimality, minimax rate, minimum distance estimation, total boundedness.

1991 Mathematics Subject Classifications: Primary 62G05.

Running Head: classifying densities

The first author's work was supported by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. The second authors' work was supported by DIGES Grant PB96-0300.

Table of Contents

§1. Introduction

Let $\mathcal{F}$ be a class of densities on the real line. Assume that the sequence of i.i.d. random variables $X_1, X_2, \ldots$ is drawn according to some density f which may, or may not belong to $\mathcal{F}$. The question is if it is possible to detect from a sample whether the underlying density is a member of the class $\mathcal{F}$. Examples of $\mathcal{F}$ include the class of all normal densities, the class of all densities bounded by 5, the class of all unimodal densities, or the class of densities supported on $[0, 5]$ which are Lipschitz with a Lipschitz constant less than 2. We investigate the discernibility of such properties.

A **classification rule** is a sequence $\{T_n\}$ of functions $T_n : \mathcal{R}^n \to \{0, 1\}$ so that upon observing the sample $X_1, \ldots, X_n$, one guesses that the unknown density f is in $\mathcal{F}$ iff $T_n(X_1, \ldots, X_n) = 1$. A class $\mathcal{F}$ is called almost surely discernible, or simply **discernible**, if there exists a classification rule such that for any density f

$$\mathrm{P}\left\{T_n(X_1, \ldots, X_n) \neq I_{\{f \in \mathcal{F}\}} \quad \text{for only finitely many } n\right\} = 1.$$

(Here I denotes the indicator function.) In other words, we require that the classification rule make the right decision eventually, almost surely, for any density. A classification rule $\{T_n\}$ with the above property is called **consistent**. Obviously, $\mathcal{F}$ is discernible if and only if its complement $\mathcal{F}^c$ is discernible.

The above definition of discernibility was introduced by Dembo and Peres (1994). More precisely, Dembo and Peres call two classes of densities $\mathcal{F}$ and $\mathcal{G}$ discernible if there exists a classification rule such that, based on an i.i.d. sample, for any $f \in \mathcal{F} \cup \mathcal{G}$ the rule makes at most finitely many mistakes almost surely. In this paper we restrict our attention to the special case $\mathcal{G} = \mathcal{F}^c$. Dembo and Peres base their definition on similar notions appearing in the early work of Hoeffding and Wolfowitz (1958) and LeCam and Schwartz (1960). Other related work is by Kulkarni and Zeitouni (1995) who studied a similar, though slightly weaker definition of discernibility. Their definition, based on previous work of Cover (1973) and Koplowitz (1977), is asymmetric. If $f \notin \mathcal{F}$ they allow the classification rule to fail for some very small subclass of densities. Under their definition, Kulkarni and Zeitouni propose a general classification procedure. This, however, cannot be used to prove discernibility under the symmetric definition studied here.

The problem we are looking at here is basically a composite hypothesis testing problem. The Dembo-Peres definition we chose to adopt is merely one of the several meaningful possibilities. In fact, the variety of similar definitions introduced in the literature often leads to seemingly contradictory results. A property of a density may be testable according to one definition and not testable according to another. One such example is unimodality of the density. One of the examples we show below is that the class of all unimodal densities is discernible. This is in contrast with a result of Donoho (1988) who, working with a stronger notion of testability, showed that unimodality is not testable. Preferring one definition over another may depend on the particular application one has in mind, or simply may be a matter of taste.

This paper points out that many common assumptions one finds nowadays in papers simply are not discernible. For example, the class of compact support densities is not discernible. Similarly, the class of densities with continuous first derivative is not discernible. Integrability conditions such as $\int f^2 < \infty$ can also not be detected in the sense used in this paper. On the other hand, discernible classes include the class of all unimodal densities, all monotone densities on the halfline, and all convex densities on a halfline. Several general theorems in this respect are provided. For example, closed classes with an L_1 minimax risk over $\mathcal{F}$ that tends to zero are discernible.

Even though we provide several explicit and non-obvious constructions of classification rules, our primary interest is in determining which classes are discernible, and we do not aim to construct hypothesis tests which are optimal in some classical sense. The reader will certainly find room for obvious improvements. Whenever possible, we try to give the simplest possible consistent classification rule.

Even though we work with densities on the real line, basically all results may be reproduced for multivariate densities in $\mathcal{R}^d$. Since the multivariate problem does not need new ideas, we stay with the simple case of $d = 1$.

The paper is organized as follows. In Section 2 simple sufficient conditions of discernibility are obtained for classes of the form $\{f : \Psi(f) \le c\}$ where Ψ is a functional and c is a known constant. Theorem 2, one of the main results of the paper, is presented in Section 3. The theorem states that closed classes (in L_1) are discernible if there exists a density estimate whose L_1 error converges uniformly in the class. Sections 4, 5, 6, and 7 contain several applications of Theorem 2. In these sections various general sufficient conditions are derived for Theorem 2. Many important specific classes are shown to be discernible. In Section 8 an ad hoc analysis is carried out to prove discernibility of certain smoothness and monotonicity classes which do not satisfy the conditions of Theorems 1 and 2. The proof of Theorem 2 is given in Section 9. In Section 10 a general sufficient condition is obtained for non-discernibility and several examples and shown. Section 11 presents further negative examples.

§2. Classes defined by functionals

In this section we derive some sufficient conditions for discernibility for classes which are defined in terms of some functional of the density. A functional Ψ assigns an (extended) real number to every density. Examples include $\Psi(f) = \int f^2$; $\Psi(f) = \int f \log f$; $\Psi(f) = \operatorname{ess\,sup} f$; or $\Psi(f) = \int (f'^2/f)$.

THEOREM 1. *Let Ψ be a functional defined for all densities, and consider a class $\mathcal{F} = \{f : \Psi(f) \le c\}$, where c is a constant. Assume that there exists an estimate $\Psi_n = \Psi_n(X_1, \ldots, X_n)$ of $\Psi(f)$ and a sequence $a_n \to 0$ such that for all densities f,*

$$\sum_{n=1}^{\infty} \mathbb{P}\{\Psi_n > \Psi(f) + a_n\} < \infty,$$

and $\Psi_n \to \Psi(f)$ almost surely. (Note: the convergence of $(\Psi_n - \Psi)_+$ must thus be uniform, but not that of $(\Psi - \Psi_n)_+$.) Then $\mathcal{F}$ is discernible.

REMARK 1. An explicit classification rule based on Theorem 1 can only be constructed if a_n is explicitly known. We will provide several examples following the proof.

Remark 2. In contrast to Theorem 1, the class $\mathcal{F} = \{f : \Psi(f) < \infty\} = \cup_{c \in \mathcal{R}} \{f : \Psi(f) \le c\}$ of densities for which $\Psi(f)$ is finite is often not discernible, even if the conditions of Theorem 1 are satisfied. Examples are presented in Sections 10 and 11.

Proof. Consider the classification rule

$$T_n(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } \Psi_n \le c + a_n, \\ 0 & \text{otherwise.} \end{cases}$$

If the common density f of the X_i's is in the class $\mathcal{F}$ then the probability of making a mistake is

$$\mathbf{P}\{T_n = 0\} = \mathbf{P}\{\Psi_n > c + a_n\} \le \mathbf{P}\{\Psi_n > \Psi(f) + a_n\},$$

which is summable by assumption, so the Borel-Cantelli lemma implies that the number of mistakes remains finite almost surely. If, on the other hand, $f \notin \mathcal{F}$, then $\Psi(f) > c$, so since $\Psi_n \to \Psi(f)$ almost surely, for sufficiently large n, $\Psi_n > c + a_n$, and therefore the number of mistakes is finite in this case as well. $\square$

Example 1: The class of densities supported in $[-c, c]$. As a first simple example, let $c > 0$ be a fixed and known constant, and consider the class $\mathcal{F}_c$ of densities supported in the interval $[-c, c]$. This class is discernible. To prove this, define $\Psi(f) = \operatorname{ess\,sup}\{|x| : f(x) > 0\}$, and $\Psi_n = \max_{i \le n} |X_i|$. The classification rule $T_n(X_1, \ldots, X_n) = I_{\Psi_n \le c}$ is easily shown to be consistent. However, ignoring the existence of this obvious classification rule, we may also apply Theorem 1. Clearly, $\Psi_n \to \Psi(f)$ almost surely. Furthermore, for any $\epsilon > 0$, $\mathbf{P}\{\Psi_n > \Psi(f) + \epsilon\} = 0$, and therefore the condition of Theorem 1 is satisfied for any positive sequence $a_n \to 0$. $\square$

Example 2: The class of densities bounded by c. Let $c > 0$, and consider the class $\mathcal{F}_c$ of all densities such that $\operatorname{ess\,sup} f \le c$. Then $\mathcal{F}_c$ is discernible. To prove this statement, we apply Theorem 1 with $\Psi(f) = \operatorname{ess\,sup} f$. The estimate of the functional $\Psi(f)$ is $\Psi_n = \Psi(f_n) = \sup_x f_n(x)$, where f_n is the kernel estimate of f defined by

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} I_{\{|x - X_i| \le h/2\}},$$

where we choose the smoothing factor h to be $h = n^{-1/3}$. First we check the property involving a_n. For any $\epsilon > 0$ and for every n,

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \mathbf{P}\{\Psi_n > \Psi(f) + \epsilon\} &= \sup_{f \in \mathcal{F}} \mathbf{P}\left\{\sup_x f_n > \operatorname{ess\,sup} f + \epsilon\right\} \\
&\le \sup_{f \in \mathcal{F}} \mathbf{P}\left\{\sup_x \left(\frac{1}{nh}\sum_{i=1}^{n} I_{\{|x - X_i| \le h/2\}} - \frac{1}{h}\int_{x-h/2}^{x+h/2} f(z)dz\right) > \epsilon\right\} \\
&\le \sup_{f \in \mathcal{F}} \left[\mathbf{P}\left\{\sup_x \left(\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}} - \int_{-\infty}^{x} f(z)dz\right) > \frac{h\epsilon}{2}\right\}\right. \\
&\quad \left. + \mathbf{P}\left\{\sup_x \left(\int_{-\infty}^{x} f(z)dz - \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}}\right) > \frac{h\epsilon}{2}\right\}\right] \\
&\le 2e^{-nh^2\epsilon^2/2} \\
&= 2e^{-n^{1/3}\epsilon^2/2},
\end{aligned}$$

where the last inequality follows from Massart's (1990) sharpened version of the Dvoretzky-Kiefer-Wolfowitz theorem. Therefore, we may take $a_n = n^{-1/7}$ in Theorem 1. To complete the proof, we need to show that for any density,

$$\sup_x f_n(x) \to \operatorname{ess\,sup} f \quad \text{almost surely.}$$

It is clear from the argument above that for any f, $\limsup_n \Psi_n \leq \Psi(f)$ almost surely. On the other hand, observe that for any f, by the Lebesque density theorem (see, e.g., Wheeden and Zygmund, 1977), for every $B < \operatorname{ess\,sup} f$ there exists an $x_0 \in \mathcal{R}$ and a $\delta_0 > 0$ such that for all $\delta < \delta_0$, $\int_{x_0-\delta/2}^{x_0+\delta/2} f(z)dz > B\delta$. Then if n is so large that $h = n^{-1/3} < \delta_0$, then for any $\epsilon > 0$,

$$\begin{aligned} \mathbf{P}\left\{f_n(x_0) < B - \epsilon\right\} &\leq \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n} I_{\{|x_0-X_i|\leq h/2\}} < \int_{x_0-h/2}^{x_0+h/2} f(z)dz - \epsilon h\right\} \\ &\leq e^{-2n\epsilon^2h^2} = e^{-2n^{1/3}\epsilon^2} \end{aligned}$$

by Hoeffding's inequality (1963). This, together with the Borel-Cantelli lemma proves that, almost surely, $\liminf_n \Psi_n \geq \Psi(f)$, concluding the proof. □

Example 3: The class of densities whose integrated square is at most c. Let $c > 0$, and define the class

$$\mathcal{F}_c = \left\{f : \int f^2 \leq c\right\} .$$

The estimator of $\Psi(f) = \int f^2$ may be $\Psi_n = \int f_n^2$ or $\Psi_n = (2/n)\sum_{i=n/2+1}^{n} f_{n/2}(X_i)$, where f_n is a kernel or histogram estimate of f. For the sake of simplicity, we take $\Psi_n = \sum_i N_i^2/(n^2h)$, where $N_i, i \in \mathcal{Z}$, are the cardinalities of the intervals $[ih, ih+h)$. Observe that changing one X_j and replacing it by X_j', causes at most two N_i's to change by one. Therefore, the change in Ψ_n is at most $2/(nh)$. By McDiarmid's inequality (McDiarmid, 1989),

$$\mathbf{P}\{|\Psi_n - \mathbf{E}\{\Psi_n\}| > \epsilon\} \leq 2\exp(-nh^2\epsilon^2/2) .$$

If $nh^2/\log n \to \infty$, the probability on the left-hand-side is summable, uniformly over all f. In particular, $\Psi_n \to \Psi(f)$ almost surely (even if $\Psi(f) = \infty$) if $\mathbf{E}\{\Psi_n\} \to \Psi(f)$. Denote $A_i = [hi, h(i+1))$, $p_i = \int_{A_i} f$. Then

$$\begin{aligned} \mathbf{E}\{\Psi_n\} &= \frac{\sum_i (np_i)^2 + np_i(1-p_i)}{n^2h} \\ &= \int g_h^2 + \frac{\sum_i p_i(1-p_i)}{nh} \\ &\qquad \text{(where } g_h(x) = p_i/h, x \in A_i \text{ is a density)} \\ &= \int g_h^2 + o(1) . \end{aligned}$$

Assume $h \to 0$ as $n \to \infty$. By the Lebesgue density theorem (Wheeden and Zygmund, 1977), $g_h \to f$ at almost all x when $h \to 0$. Thus, by Fatou's lemma, $\liminf_{n\to\infty} \mathbf{E}\{\Psi_n\} \geq \int \liminf_{h\to 0} g_h^2 = \int f^2$. This remains valid even if $\int f^2 = \infty$. Furthermore, by Jensen's inequality, $\int g_h^2 \leq \int f^2$. Collecting all this shows that for all f, $\Psi_n \to \Psi(f)$ almost surely if $h \to 0$, $nh^2/\log n \to \infty$. Furthermore, if we take $a_n = 1/n^{1/5}$, then for all f,

$$\begin{aligned} \mathbf{P}\{\Psi_n > \Psi(f) + a_n\} &\leq \mathbf{P}\{\Psi_n - \mathbf{E}\{\Psi_n\} > a_n/2\} \quad (\text{if } 1/nh < a_n/2) \\ &\leq 2\exp(-nh^2/8n^{2/5}) \end{aligned}$$

and this is summable in n if we take, say, $h = n^{-1/5}$. By Theorem 1, the class of densities is thus discernible by the classification rule $\Psi_n \leq c + 1/n^{1/5}$. □

Example 4: The class of densities f with $\int f^r \leq c$, where $r > 1$ is a fixed constant. The details for this extend those of the previous example, and are left to the reader. □

§3. Classes with vanishing minimax risk

Next we derive another general sufficient condition for discernibility. We recall that a density estimate f_n is a real-valued measurable function of $x \in \mathcal{R}$ and the data $X_1, \ldots, X_n$:

$$f_n(x) = f_n(x, X_1, \ldots, X_n).$$

f_n is called **strongly universally consistent** if for every density f,

$$\lim_{n\to\infty} \int |f_n - f| = 0 \quad \text{almost surely.}$$

Note that strongly universal consistent estimates exist (see, e.g., Devroye and Györfi, 1985). A density estimate f_n is **uniformly convergent on** $\mathcal{F}$ if

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \mathrm{E} \int |f_n - f| = 0 \ .$$

Also, we say that $\mathcal{F}$ is closed (in the L_1 space of densities) if for any density $g \notin \mathcal{F}$ there exists an $\epsilon > 0$ and an open ball $B_{g,\epsilon} = \{f : \int |f - g| < \epsilon\}$ such that $B_{g,\epsilon} \cap \mathcal{F} = \emptyset$.

Theorem 2. *Let $\mathcal{F}$ be a closed class of densities. If there exists a density estimate f_n that is uniformly convergent on $\mathcal{F}$, then $\mathcal{F}$ is discernible. And any closed subclass $\mathcal{G} \subseteq \mathcal{F}$ is discernible.*

The proof of Theorem 2 is postponed to Section 9.

As the following Lemma shows, apart from the closedness of the class, we require the convergence to zero of the minimax expected L_1 error of the class.

Lemma 1. *The following are equivalent:*

A.

$$\lim_{n\to\infty} \inf_{f_n} \sup_{f\in\mathcal{F}} \mathrm{E} \int |f_n - f| = 0 \ .$$

B. There exists a density estimate f_n such that

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \mathrm{E} \int |f_n - f| = 0 \ .$$

Proof. B implies A. A implies B because A says that for each k, we may find density estimates $f_{n,k}$ and numbers n_k such that for $n \geq n_k$, $\sup_{f\in\mathcal{F}} \mathrm{E} \int |f_{n,k} - f| < 1/k$. Without loss of generality, $n_k \uparrow$. Define $g_n = f_{n,k}$, $n_k \leq n < n_{k+1}$, and use g_n in part B. □

It is well-known (Devroye, 1983) that very large classes do not satisfy this latter condition. However many positive examples are also known. The theorem also points out that we do not have to explicitly construct f_n—a proof of existence suffices. Note however that the actual density estimate used to test membership in the class $\mathcal{F}$ may be very different from the uniformly consistent f_n. A construction follows from the proof below. The reason we need a special construction is that f_n may not be stable or concentrated enough to provide good almost sure behavior over an entire sequence.

Example 5: Finite classes of densities. As the kernel estimate is universally consistent if we take the smoothing factor h such that $h \to 0$ and $nh \to \infty$ ($nh^d \to \infty$ in R^d), it is clear from Theorem 2 that we can always construct a consistent classification rule for the class $\mathcal{F} = \{f\}$. In fact, any finite class of densities is discernible. $\Box$

§4. Discernibility via kernel estimates

The next lemma states that to verify the minimax condition, it suffices to establish it for a kernel estimate in which the bandwidth may depend upon the (unknown) density. It removes the burden of construction of minimax optimal or near-optimal densities. Recall that the kernel estimate of a density is

$$f_{nh}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $h > 0$ is a **bandwidth** and $K : \mathcal{R} \to \mathcal{R}$ is a kernel function usually chosen to satisfy $\int K = 1$. Note that all kernels used in practice satisfy the condition of the following Lemma.

LEMMA 2. *Assume that g_n is a kernel estimate with kernel K of polynomial kernel complexity (Devroye and Lugosi, 1997), and with bandwidth $h_n = h_n(f)$. Then, if*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \mathbf{E} \int |g_n - f| = 0 \ ,$$

it follows that there exists a density estimate f_n such that

$$\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \mathbf{E} \int |f_n - f| = 0 \ ,$$

PROOF. This follows from the inequalities of Devroye and Lugosi (1996, 1997), where a data-dependent bandwidth is constructed guaranteeing that the kernel estimate f_n with that bandwidth satisfies $\mathbf{E} \int |f_n - f| \leq C \inf_h \mathbf{E} \int |f_{nh} - f| + D\sqrt{\log n/n}$ for universal constants D and C not depending upon f, and all n. Here f_{nh} is the kernel estimate with deterministic bandwidth h. $\Box$

The lemma above allows us to further deduce sufficient conditions for our classes $\mathcal{F}$, while avoiding explicit constructions of density estimates for them.

LEMMA 3. *If $\mathcal{F}$ and $\mathcal{G}$ are disjoint classes of densities that $\mathcal{G}$ and $\mathcal{F} \cup \mathcal{G}$ are both discernible, then so is $\mathcal{F}$. If $\mathcal{F}$ and $\mathcal{G}$ are discernible, then so are $\mathcal{F} \cup \mathcal{G}$ and $\mathcal{F} \cap \mathcal{G}$.*

PROOF. We run a classification rule for $\mathcal{G}$, which based on a sample of size n yields the decision Y_n. Then we run a classification rule on the same sample for membership in $\mathcal{F} \cup \mathcal{G}$, and call the decision Z_n. We decide that the unknown density is in $\mathcal{F}$ when $Z_n = 1$ and $Y_n = 0$. This makes a finite number of errors with probability one. The second statement does not require disjointness: we decide $\mathcal{F} \cup \mathcal{G}$ if $W_n + Y_n \geq 1$, where W_n is the decision for $\mathcal{F}$. Again, only a finite number of errors are made with probabilty one. Finally we decide $\mathcal{F} \cap \mathcal{G}$ if $W_n Y_n = 1$. $\Box$

EXAMPLE 6: SCALE/TRANSLATION CLASSES. As a first example, let $\mathcal{F}$ be any subset of the scale/translation class of densities $f((\cdot - \mu)/\sigma)/\sigma$, where f is a fixed density, $\mu \in \mathbf{R}$ and $\sigma > 0$. Then, in any kernel estimate, take $h_n = \sigma/\sqrt{n}$, and note that with this choice, the expected L_1 error is the same for all μ and σ. Therefore, by Lemma 2, the L_1 minimax risk tends to zero for this scale/translation class and hence for $\mathcal{F}$. Therefore, $\mathcal{F}$ is discernible whenever it is closed. In particular, we can test whether a normal density has a rational $\sigma = p/q$, with $p > 0$ and $0 < q \leq 500$. And normality is discernible, as is any scale/translation class (by closedness). $\square$

Example 7: Nonlinear transformation classes. As a second example, let $\mathcal{F}$ be the class of all densities for random variables $T(X)$, where $X > 0$ has density f, and $T(x) = (x^a - 1)/a$, $a > 0$ or $T(x) = \log(x)$ (case $a = 0$) (the Box-Cox transformations). Then $\mathcal{F}$ is discernible. To see this, transform the data by $T^{-1}(\cdot)$, and use a kernel estimate for f with bandwidth $h_n = 1/\sqrt{n}$. Then note that the inequality in the proof of Lemma 2 has been extended to include the joint data-based choice of h and the Box-Cox parameter a (Devroye, Lugosi and Udina, 2000). Therefore, using the fact that L_1 errors are invariant under monotone transformations of the input, we conclude there exists a density estimate for which the expected L_1 error tends to zero uniformly over all values of a. As the density of $T(X)$ has no limit in the class of densities when $a \to \infty$, we see that $\mathcal{F}$ is closed, and conclude that $\mathcal{F}$ is discernible. $\square$

Example 8: Parametric classes. Consider a countable class $\mathcal{F} = \{\phi_k, 1 \leq k \leq \infty\}$ of densities ϕ_∞ and ϕ_k $(k < \infty)$, with $\phi_k \to \phi_\infty$, where convergence is meant in the L_1 sense. This class is closed in the collection of all densities. Also, its L_1 minimax error tends to zero. To see this, consider the kernel estimate with bandwidth $h_n = 1/\sqrt{n}$. Then the expected L_1 error is easily bounded as follows: let f and g be two densities, and let f_n and g_n be two kernel estimates based upon two samples of size n, both using the same kernel and bandwidth. Then there exists a coupling of the samples such that

$$\mathrm{E}\int |f_n - g_n| \leq \left(1 + \int |K|\right) \int |f - g|$$

(Devroye, 1985). We take K such that $\int |K| = 1$. So, given $\epsilon > 0$, let K_ϵ be the collection of indices $k < \infty$ such that $\int |\phi_k - \phi_\infty| > \epsilon$. Then, with f_{nk} denoting the kernel estimate mentioned above (with n for sample size, and k for ϕ_k), we have, letting g_n denote the kernel estimate (with the same kernel and bandwidth as f_{nk}) based on a sample of size n from ϕ_∞,

$$\begin{aligned}
&\sup_k \mathrm{E}\int |f_{nk} - \phi_k| \\
&\leq \max_{k \in K_\epsilon} \mathrm{E}\int |f_{nk} - \phi_k| + \sup_{k \notin K_\epsilon} \left(\mathrm{E}\int |f_{nk} - g_n| + \mathrm{E}\int |g_n - \phi| + \int |\phi_k - \phi|\right) \\
&\leq \max_{k \in K_\epsilon} \mathrm{E}\int |f_{nk} - \phi_k| + \mathrm{E}\int |g_n - \phi_\infty| + 3 \sup_{k \notin K_\epsilon} \int |\phi_k - \phi_\infty| \\
&\leq o(1) + 3\epsilon\ .
\end{aligned}$$

Thus, $\mathcal{F}$ is discernible. (This fact may also be derived as a consequence of Theorem 3 stated below.) By arguments as above, the class $\mathcal{G}$ consisting of all scaled/or translated densities from $\mathcal{F}$ is also discernible. To see why this example is powerful, let X_k denote a random variable with the gamma density $x^{k-1}e^{-x}/\Gamma(k)$, $x > 0$, $k > 0$, k integer. Let $Y_k = (X_k - k)/\sqrt{k}$ be the normalized random variable. Then, as is well-known, the density of Y_k tends to the normal $(0, 1)$ density in L_1 as $k \to \infty$. As argued above, the class $\mathcal{F}$ of all the densities of Y_k, merged with the standard normal density, is discernible. If we consider the scale/translation enlargement $\mathcal{G}$ of $\mathcal{F}$, we obtain all gamma densities, possibly linearly transformed, with integer shape parameter plus all normal densities. It too is discernible. Finally, a small additional argument shows that the shape parameter does not have to be restricted at all. Thus, parametric classes in general, when merged with their limit densities, are in most instances discernible. When the limits are not included, the question is different, but Lemma 3 provides help. To illustrate this, let $\mathcal{F}$ be the class of all scaled and translated gamma densities, and $\mathcal{G}$ be the class of all normal densities. Then, as argued above, both $\mathcal{F} \cup \mathcal{G}$ and $\mathcal{G}$ are discernible, so $\mathcal{F}$ is. Lemma 3 is thus useful for the removal of limit classes of parametric collections. Finally, let $\mathcal{F}$ be the class of all gamma densities, not translated or scaled. This class does not contain any densities as limits, and is thus closed. Hence it is discernible. □

§5. Discernibility via total boundedness

A class $\mathcal{F}$ of densities is said to be totally bounded if for every $\epsilon > 0$, there exists a finite number N_ϵ of densities $f_k, 1 \le k \le N_\epsilon$ such that the L_1 balls of radius ϵ centered at the f_k's cover $\mathcal{F}$. (Note that in this definition, it is irrelevant whether we additionally ask that the f_k's belong to $\mathcal{F}$: both definitions would be equivalent.) That is, for every $f \in \mathcal{F}$, there exists $k \le N_\epsilon$ such that $\int |f - f_k| \le \epsilon$. The smallest possible value of $\log N_\epsilon$ for $\mathcal{F}$ is called the Kolmogorov entropy of $\mathcal{F}$. Yatracos (1985) (see also Devroye, 1987, p. 90) has constructed a minimum distance estimate f_n with the property

$$\sup_{f\in\mathcal{F}} \mathbf{E}\left\{\int |f_n - f|\right\} \le 5\epsilon + \frac{4+\sqrt{128N_\epsilon}}{\sqrt{2n}} \ .$$

Define $\epsilon_k = 1/k + \inf\{\epsilon > 0 : 128 \log N_\epsilon < \sqrt{k}\}$. By total boundedness, $\epsilon_k \to 0$ as $k \to \infty$. Thus,

$$\sup_{f\in\mathcal{F}} \mathbf{E}\left\{\int |f_n - f|\right\} \le 5\epsilon_n + \frac{4+\sqrt{128N_{\epsilon_n}}}{\sqrt{2n}} \le o(1) + \frac{4}{\sqrt{2n}} + \frac{8}{n^{1/4}} \ .$$

Thus, applying Theorem 2, we immediately have

THEOREM 3. *Let $\mathcal{F}$ be a closed totally bounded class of densities. Then $\mathcal{F}$ is discernible.*

EXAMPLE 9: BOUNDED UNIMODAL DENSITIES. Let $\mathcal{F}$ be any closed class of unimodal densities bounded by B with support on $[0,1]$. Then $\mathcal{F}$ is discernible. In particular, the class of all concave densities with support on $[0,1]$ is discernible.

EXAMPLE 10: LIPSCHITZ DENSITIES. Let $\mathcal{F}$ be the class of Lipschitz densities on $[0,1]$ with Lipschitz constant not exceeding C. Then $\mathcal{F}$ or any closed subclass of it is discernible.

EXAMPLE 11: FINITE MIXTURES. Let $\mathcal{F}$ be the class of convex mixtures of k fixed densities. Clearly, this class is closed. Also, by creating a finite grid for the possible convex weights, it is trivial to see that this class is totally bounded. Thus, $\mathcal{F}$ is discernible.

EXAMPLE 12: UNIFORM MODULUS OF CONTINUITY. Let $\mathcal{F}$ be a closed class of densities on $[0,1]$ with uniformly bounded modulus of continuity: for all $\delta > 0$,

$$\sup_{f\in\mathcal{F}} \sup_{x,y:\|y-x\|\le\delta} |f(y) - f(x)| < \infty \ .$$

Then $\mathcal{F}$ is totally bounded (Lorentz, 1966; see also Devroye, 1987, p. 98) and thus discernible.

§6. Discernibility via Yatracos' minimum distance estimates

Next we give another simple sufficient condition for discernibility via Theorem 2. Recall that given a class of sets $\mathcal{A}$, the VC **dimension** V of $\mathcal{A}$ is defined as the largest positive integer k for which there exist k points $x_1, \ldots, x_k$ such that

$$|\{A \cap \{x_1, \ldots, x_k\} : A \in \mathcal{A}\}| = 2^k.$$

If there is no such largest k, then we say that $V = \infty$. If $V < \infty$, then $\mathcal{A}$ is called a VC **class**.

THEOREM 4. *Let $\mathcal{F}$ be a closed class of densities, and assume that the class of sets*

$$\mathcal{A} = \{x : f(x) > g(x); f, g \in \mathcal{F}\}$$

is a VC *class. Then $\mathcal{F}$ is discernible.*

PROOF. By Theorem 2, it suffices to prove that there exists a density estimate f_n which is uniformly convergent in $\mathcal{F}$. Consider now the minimum distance estimate proposed by Yatracos (1988):

$$f_n = \arg\min_{f \in \mathcal{F}} \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right|,$$

where $\mu_n(A) = (1/n) \sum_{i=1}^n I_A(X_i)$ is the empirical measure of A based on the random sample. Yatracos showed (see also Devroye, Györfi, and Lugosi, 1996, p. 278) that

$$\sup_{f \in \mathcal{F}} \mathbf{E} \int |f_n - f| \le 4\, \mathbf{E} \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right|.$$

The well-known Vapnik-Chervonenkis inequality (Vapnik and Chervonenkis, 1971) implies that

$$\mathbf{E} \left\{ \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right| \right\} \le 8 \sqrt{\frac{V \log n + 3}{2n}}$$

(see also Devroye, Györfi, and Lugosi, 1996), which finishes the proof. □

EXAMPLE 13: EXPONENTIAL MIXTURES. Let k be a fixed positive integer, and consider the class $\mathcal{F}$ of all mixtures of k exponential densities (that is, translations and scales of $e^{-x}, x \ge 0$. Then Theorem 4 may be applied to show that $\mathcal{F}$ is discernible. It suffices to show that the class

$$\mathcal{A} = \{x : f(x) > g(x); f, g \in \mathcal{F}\}$$

is a VC class. A member set in this class is thus of the form

$$\left\{ x : \sum_{i=1}^{2k} a_i e^{-b_i x} I_{[x > c_i]} > 0 \right\}$$

where $a_i, c_i \in \mathbf{R}$ and $b_i > 0$ are free parameters. We claim that each set in this class is the union of at most $\ell \stackrel{\text{def}}{=} (2k+1)(k+1)$ intervals. Since the class of unions of ℓ intervals is a VC class, we are done. Therefore, the classification rule described above may be used to test whether a density is a k-mixture of exponentials. We may similarly test for k-mixtures of normals and indeed many other parametric families. The interval-counting argument is as follows: clearly, we have at most $k+1$ intervals defined by the thresholds c_i. It suffices to show that on each of these intervals, a set of the form $\{x : \sum_{i=1}^{2k} a_i e^{-b_i x} > 0\}$ defines at most $2k+1$ intervals. But this is well-known (Lemma 25.2 of Devroye, Györfi and Lugosi, 1996). Hence the claim.

§7. Smoothness and monotonicity classes

In this section, we remove the burden of checking the conditions of Theorem 2, and provide easy-to-verify sufficient conditions for classes of densities to be discernible. We begin with a subclass $\mathcal{F}$ of the monotone densities f on $[0, \infty)$ with finite value $f(0)$.

THEOREM 5. *Let $\mathcal{F}$ be a closed subclass of the monotone densities f with finite modal value $m(f)$. Assume that*

$$\sup_{f\in\mathcal{F}} \left(\int \sqrt{f}\right)^{\frac{2}{3}} (m(f))^{\frac{1}{3}} < \infty \ .$$

Then $\mathcal{F}$ is discernible.

PROOF. Theorem 5 follows from Theorem 2 and Lemmas 1 and 2 if, for $f \in \mathcal{F}$, we can establish uniform L_1 error bounds for the kernel estimate f_{nh} with uniform kernel K on $[-1,1]$. To this end, we use the typical argument (see Devroye and Györfi, 1985). From the proof below, it will become apparent that the position of the mode is unimportant, and thus we may assume without loss of generality that the mode occurs at zero, and thus that $m(f) = f(0)$. Let $*$ be the convolution operator.

$$\begin{aligned}
\mathbf{E}\int |f_{nh} - f| &\le \int |\mathbf{E}f_{nh} - f| + \int \sqrt{\mathbf{E}(f_{nh} - \mathbf{E}f_{nh})^2} \\
&= \int |f * K_h - f| + \int \sqrt{(1/n)\mathbf{V}(K_h(x - X_1))} \\
&\le \int |f * K_h - f| + \int \sqrt{(1/n)\mathbf{E}\{K_h^2(x - X_1)\}} \ .
\end{aligned}$$

For $x > h$, observe that $|f*K_h(x) - f(x)| \le f(x-h) - f(x+h)$, and that for $x < h$, $|f*K_h(x) - f(x)| \le f(0)$. Furthermore, $\mathbf{E}\{K_h^2(x - X_1)\} \le f(\max(0, x-h))/(2h)$. Using these estimates, we obtain

$$\begin{aligned}
\mathbf{E}\int |f_{nh} - f| &\le hf(0) + \int_h^{3h} f + h\sqrt{\frac{f(0)}{2nh}} + \frac{\int\sqrt{f}}{\sqrt{2nh}} \\
&\le 3hf(0) + h\sqrt{\frac{f(0)}{2nh}} + \frac{\int\sqrt{f}}{\sqrt{2nh}} \\
&= \frac{4(\int\sqrt{f})^{\frac{2}{3}}(f(0))^{\frac{1}{3}}}{(2n)^{\frac{1}{3}}} + \frac{(\int\sqrt{f})^{\frac{1}{3}}(f(0))^{\frac{1}{6}}}{(2n)^{\frac{2}{3}}}
\end{aligned}$$

when we take $h^{3/2} = \int\sqrt{f}/(f(0)\sqrt{2n})$. Therefore, we have uniform convergence to zero whenever

$$\sup_{f\in\mathcal{F}} \left(\int \sqrt{f}\right)^{\frac{2}{3}} (f(0))^{\frac{1}{3}} < \infty \ .$$

By Lemma 2 and Theorem 2, Theorem 5 follows. □

Let us extend the previous Theorem to include densities of bounded variation. We recall that a density f is of bounded variation if we may find an increasing function f_1 and a decreasing function f_2 such that $f = f_1 + f_2$, such that the total variation

$$\mathcal{V}(f) \stackrel{\text{def}}{=} \inf_{f_1,f_2:f=f_1+f_2,f_1\uparrow,f_2\downarrow} \sup_{y>x}(f_1(y) - f_1(x)) + \sup_{y>x}(f_2(x) - f_2(y)) < \infty \ .$$

THEOREM 6. *Let $\mathcal{F}$ be a closed subclass of the densities of bounded variation, and assume that*

$$\sup_{f\in\mathcal{F}} (\mathcal{V}(f))^{\frac{1}{3}} \left(\int \sqrt{f}\right)^{\frac{2}{3}} < \infty \ .$$

Then $\mathcal{F}$ is discernible.

PROOF. We argue as in the proof of Theorem 5, and apply Lemma 2 and Theorem 2. Let f_{nh} be the kernel estimate. If $f = f_1 + f_2$ is the bounded variation decomposition of f, with $f_1 \uparrow$ and $f_2 \downarrow$, and K is the uniform kernel on $[-1,1]$, then

$$|f * K_h - f| \leq f_1(x+h) - f_1(x) + f_2(x-h) - f_2(x)$$

and

$$|f * K_h - f| \leq 2h\mathcal{V}(f) \ .$$

Also,

$$\int \sqrt{\frac{\mathrm{E}\{K_h^2(x-X_1)\}}{n}} \leq \int \sqrt{\frac{f(x+h)+f(x-h)}{2hn}}\, dx \leq \sqrt{\frac{2}{nh}} \int \sqrt{f}$$

so that

$$\mathrm{E}\left\{|f_{nh} - f|\right\} \leq 2h\mathcal{V}(f) + \sqrt{\frac{2}{nh}} \int \sqrt{f} = \frac{4}{(2n)^{1/3}} \left(\int \sqrt{f} \right)^{\frac{2}{3}} (\mathcal{V}(f))^{\frac{1}{3}}$$

if we take $h^{3/2} = \int \sqrt{f} / (\sqrt{2n}\mathcal{V}(f))$. □

EXAMPLE 14: LOG-CONCAVE DENSITIES. A very important subclass of the densities is the class of log-concave densities: $\log f$ is concave. We claim that any closed subclass of the log-concave densities is discernible. Consider that this class includes all beta densities with both parameters greater than or equal to 1, all gamma densities with shape parameter greater than or equal to one, all exponential power distributions with shape parameter at least one, the normal densities, and a host of other densities. It is known that these densities are unimodal and that if the mode occurs at z, a rescaling of the random variable to place the mode at zero with modal value one results in a density g with $g(u) \leq e^{1-|u|}$ (Devroye, 1984). As the product in Theorem 5 is scale and translation invariant, and the log-concave inequality is absolute, we see that, except for the monotonicity, all conditions of Theorem 5 are fullfilled. It is left as a trivial exercise to extend Theorem 5 to this class of densities. Thus, we have a very simple condition for establishing the discernibility of large subclasses of the famous parametric families. And as the class of log-concave densities is closed, we can indeed test log-concavity. □

EXAMPLE 15: CONCAVE DENSITIES. Consider the class of all concave densities on their support. Clearly, this class is closed. Furthermore, if the mode is forced to be at zero and of modal value one, then any density in this class is bounded by one, and of support contained in $[-2,2]$. By an argument as for the log-concave densities, this class is discernible (regardless of the support!), and indeed any closed subclass of it is discernible as well. □

One could refine the bounds of the proof of Theorem 5 and indeed use higher-order kernels to obtain results for subclasses related to Akhiezer classes of densities. In this manner, we may deal with convex subclasses of the monotone densities, and classes for which the r-th derivative $f^{(r)}$ is monotone.

EXAMPLE 16: LIPSCHITZ DENSITIES. A class $\mathcal{F}$ of Lipschitz (1) densities with Lipschitz constant $C(f) < \infty$ and support $s(f)$ is discernible if $\mathcal{F}$ is closed and

$$\sup_{f \in \mathcal{F}} C(f) s^2(f) < \infty \ .$$

To see this, apply Theorem 6, and note that $\mathcal{V}(f) \leq C(f)s(f)$, and $\int \sqrt{f} \leq \sqrt{s(f) \int f} = \sqrt{s(f)}$. □

EXAMPLE 17: UNIMODAL DENSITIES. Let $\mathcal{F}$ be a class of unimodal densities with modal value $m(f)$ and variance $\sigma^2(f)$. Then this class is discernible if $\mathcal{F}$ is closed and

$$\sup_{f \in \mathcal{F}} m(f)(1 + \sigma^2(f)) < \infty \ .$$

To see this, note that $\mathcal{V}(f) \le 2m(f)$ and assuming without loss of generality that the mean is at the origin,

$$\int \sqrt{f} = \int \frac{\sqrt{(1+x^2)f}}{\sqrt{1+x^2}} \le \sqrt{\int (1+x^2)f} \sqrt{\int \frac{1}{1+x^2}} = \sqrt{\pi(1+\sigma^2(f))} \ .$$

Examples of such classes include the gamma densities with shape parameter 1 or larger, the symmetric beta densities with shape parameter 1 or larger, and the unimodal densities $f \le A/(1+x^4)$, with $A < \infty$ fixed and given. □

§8. Ad hoc analysis: unimodality and convexity

Surprisingly, there are classes whose minimax risk and whose complement's minimax risk does not tend to zero, and yet they are discernible. We have already encountered such classes in Section 2. In this section we construct explicit classification rules to prove that the following classes are discernible: the monotone densities on $[0,\infty)$, the monotone densities on an interval of the real line, the unimodal densities with mode at 0, and the unimodal densities. To keep the material limited, we will only provide an explicit proof for the class of monotone densities on $[0,\infty)$. We recall from Devroye (1983) that the minimax risk of this class does not tend to zero. To construct an explicit ad hoc classification rule, consider a histogram estimate with bins $[0,h), [h,2h), [2h,3h), \ldots$. Let N_i denote the number of data points in $[ih,(i+1)h)$. The classification rule is the following: take $h = 1/n^{1/7}$. Then

$$\begin{aligned}&\text{decide non-monotone if } \max_{i\ge 0}(N_{i+1}-N_i) > n^{2/3} \ ;\\ &\text{decide monotone otherwise.}\end{aligned}$$

THEOREM 7. *The classification rule above makes almost surely a finite number of errors for the membership in the class of monotonically decreasing densities on* $[0,\infty)$.

PROOF. First, assume that f is indeed monotone on $[0,\infty)$. We bound $\mathbf{P}\{N_{i+1}-N_i > t\}$ by standard methods: For each $i \ge 0$, define $p_{ni} = \mathbf{P}\{X_1 \in [ih,(i+1)h)\}$. Introduce N_i' such that (N_i', N_i) are two components of a multinomial random vector with total count n and success probabilities p_{ni} each. Then

$$\begin{aligned}\mathbf{P}\{N_{i+1}-N_i > t\} &\le \mathbf{P}\{N_i' - N_i > t\}\\ &\le \mathbf{P}\{N_i' - np_{ni} > t/2\} + \mathbf{P}\{N_i - np_{ni} < -t/2\}\\ &= \mathbf{P}\{|N_i - np_{ni}| > t/2\}\\ &\le 2e^{-\frac{t^2}{2n}} \ ,\end{aligned}$$

by Hoeffding's inequality (Hoeffding, 1963). Thus,

$$\begin{aligned}&\mathbf{P}\left\{\max_{i\ge 0}(N_{i+1}-N_i) > n^{2/3}\right\}\\ &\le \mathbf{P}\left\{\max_{i\ge n} N_i > n^{2/3}\right\} + 2ne^{-n^{1/3}/2}\\ &\le \sum_{i\ge n}\mathbf{P}\left\{N_i > n^{2/3}\right\} + 2ne^{-n^{1/3}/2}\\ &\le \sum_{i\ge n}\mathbf{P}\left\{\text{binomial}(n,1/(i+1)) > n^{2/3}\right\} + 2ne^{-n^{1/3}/2}\end{aligned}$$

$$
\begin{aligned}
&\quad\text{(because } p_{ni} \le 1/(i+1) \text{ by monotonicity)}\\
&\le \sum_{i\ge n} \left(\frac{n}{(i+1)n^{2/3}}\right)^{n^{2/3}} \cdot \frac{i+1}{i} + 2ne^{-n^{1/3}/2}\\
&\quad\text{(because } \mathbf{P}\{\text{binomial}(n,p)\ge t\} \le \tbinom{n}{t}p^t/(1-p) \le (np/t)^t/(1-p),\ t \text{ integer)}\\
&= O\left(e^{-n^{1/3}/3}\right) .
\end{aligned}
$$

In the above chain, we assumed without loss of generality that $n^{2/3}$ is integer-valued. As this is summable in n, the Borel-Cantelli lemma implies that we will make finitely many errors almost surely.

Next assume that f is not monotone. Thus, there exist Lebesgue points $0 < y < z$ with the property that $f(y) = f(z) - 3\delta$ for some $\delta > 0$. As these are Lebesgue points, we know that there exists an $\epsilon > 0$ such that for any interval I of length $|I| < \epsilon$ containing y, $\int_I f(x)dx < |I|(f(y)+\delta)$, and similarly for z: $\int_I f(x)dx > |I|(f(z)-\delta)$. The number of intervals $[ih,(i+1)h)$ separating z from y is at most $2+(z-y)/h$. Let $h < \epsilon$, and let i range over the intervals covering $[y,z]$. The difference between the last p_{ni} and the first p_{ni} is at least δh, so that the maximal differential $p_{n,i+1} - p_{n,i}$ for i and $i+1$ both among the given intervals is at least $\delta h/(2+(z-y)/h) \ge \delta h^2/2(z-y) \stackrel{\text{def}}{=} 3ch^2$ for n large enough (and thus h small enough). Let j be an index for which $p_{n,j+1} - p_{n,j} \ge 3ch^2$. If we decide that f is monotone, then we have $N_{j+1} - N_j \le n^{2/3}$, however. Thus, the probability of erring is not more than

$$
\begin{aligned}
\mathbf{P}\{N_{j+1} - N_j \le n^{2/3}\} &\le \mathbf{P}\{N_{j+1} - np_{n,j+1} - (N_j - np_{nj}) \le n^{2/3} - 3cnh^2\}\\
&= \mathbf{P}\{N_{j+1} - np_{n,j+1} - (N_j - np_{nj}) \le n^{2/3} - 3cn^{5/7}\}\\
&\le \mathbf{P}\{N_{j+1} - np_{n,j+1} - (N_j - np_{nj}) \le -2cn^{5/7}\}\\
&\qquad\text{(for } n \text{ large enough)}\\
&\le \mathbf{P}\{N_{j+1} - np_{n,j+1} < -cn^{5/7}\} + \mathbf{P}\{N_j - np_{nj} > cn^{5/7}\}\\
&\le 2e^{-2c^2n^{3/7}}
\end{aligned}
$$

by Hoeffding's inequality. By the Borel-Cantelli lemma, we once again make finitely many errors almost surely. □

For monotonicity on any right-infinite interval, we apply the same classification rule, starting with the second occupied interval from the left. For unimodality, we find the interval of maximal cardinality, and apply monotonicity tests of opposite polarity on both sides of that maximal interval. The details are uninteresting. It is also worth noting that we can test for convexity. In particular, if i and j are the indices of the leftmost and rightmost intervals in a grid of intervals of width h that are occupied, then the classification rule with $h = 1/n^{1/7}$ that decides against convexity if $\max_{i<k<j-2}(2N_{k+1} - N_{k+2} - N_k) > n^{2/3}$ errs finitely often with probability one, for any density. In fact, one can in this manner test membership in any Akhiezer class.

§9. Proof of Theorem 2

The proof of the theorem is split into two lemmas, the first of which is directly applicable to prove discernibility in many cases.

LEMMA 4. *Let $\mathcal{F}$ be a closed class of densities. If there exists a sequence of positive numbers $b_n \to 0$ and a density estimate g_n such that g_n is strongly universally consistent, moreover for every $f \in \mathcal{F}$*

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{ \int |g_n - f| > b_n \right\} < \infty,$$

then $\mathcal{F}$ is discernible.

PROOF. Consider the classification rule

$$T_n(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } \inf_{f\in\mathcal{F}} \int |g_n - f| \le b_n, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, based on the data $X_1, \ldots, X_n$, we compute the density estimate g_n, and project it in the L_1 distance on the class $\mathcal{F}$. If the distance between g_n and the class is greater than b_n, then we say that the unknown density is not in the class. Now it is easy to see that this classification rule is consistent, since if the X_i's are drawn from a density $f \in \mathcal{F}$ then

$$\mathbf{P}\{T_n \text{ fails}\} = \mathbf{P}\left\{ \inf_{g\in\mathcal{F}} \int |g_n - g| > b_n \right\} \le \mathbf{P}\left\{ \int |g_n - f| > b_n \right\}.$$

By assumption these probabilities are summable over $n = 1, 2, \ldots$, so the Borel-Cantelli lemma implies that, almost surely, T_n fails at most for finitely many n.

On the other hand, if $f \notin \mathcal{F}$, then by closedness of the class $\mathcal{F}$, there exists an $\epsilon > 0$ such that $\inf_{g\in\mathcal{F}} \int |f - g| > \epsilon$. However, since g_n is strongly universally consistent and $b_n \to 0$, eventually, almost surely, $\inf_{f\in\mathcal{F}} \int |g_n - f| > \epsilon/2 > b_n$, and therefore the classification rule does not fail. □

In many cases there exists a simple estimate which satisfies the condition of Lemma 4. For example, if $\mathcal{F}$ is the class of all densities supported on $[a, b]$ such that $\operatorname{ess\,sup}_{x,y} |f(x) - f(y)|/|x - y| \le c$, then the kernel estimate with an appropriate non-data-dependent bandwidth will do. However, even simple cases as the class of all normal densities, any such kernel estimate fails to provide a uniform rate of convergence within $\mathcal{F}$. In this specific case it is possible to define a data-dependent bandwidth with the desired property (see Devroye, 1989). However, to finish the proof of Theorem 2, we need a universal construction, provided in the lemma below:

LEMMA 5. *If for some sequence $a_n \to 0$ there exists a density estimate f_n with $\sup_{f\in\mathcal{F}} \mathbf{E} \int |f_n - f| \le a_n$, then there exists another sequence $b_n \to 0$ and a density estimate g_n such that g_n is strongly universally consistent, and for all $f \in \mathcal{F}$*

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{ \int |g_n - f| > b_n \right\} < \infty.$$

REMARK 2. The proof below shows that one may always take $b_n = 3a_{\lfloor (n/2)^{1/3} \rfloor} + 3(n/2)^{-1/12} + 4n^{-1/3}$, so b_n only depends on a_n but not on the class $\mathcal{F}$. However, in most cases this is a suboptimal choice. The estimate g_n defined below is merely a part of a general proof of existence. In most concrete cases much superior estimates exist. Also, in fact, we show that $\sup_{f\in\mathcal{F}} \sum_{n=1}^{\infty} \mathbf{P}\left\{ \int |g_n - f| > b_n \right\} < \infty$.

Proof. First, we "stabilize" f_n to make sure that the L_1 error $\int |f_n - f|$ is always concentrated around its mean, and then combine it with a consistent estimate to achieve strong universal consistency. We define the stabilized density estimate $\widehat{f}_n$ as follows:

$$\widehat{f}_n(x, X_1, \dots, X_n) = \frac{1}{\lfloor \frac{n}{N} \rfloor} \sum_{k=1}^{\lfloor \frac{n}{N} \rfloor} f_N(x, X_{(k-1)N+1}, \dots, X_{kN}),$$

where $N = \lfloor n^{1/3} \rfloor$. In other words, we chop up the data into about $n^{2/3}$ equal blocks, construct the estime on all blocks, and take their average. McDiarmid's inequality (1989) assures that if by changing the value of one data point but leaving all others intact the L_1 error does not change by much, then it is close to its mean with large probability. In our case,

$$\begin{aligned}
&\left| \int |\widehat{f}_n(x, X_1, \dots, X_n) - f(x)| dx - \int |\widehat{f}_n(x, X_1, \dots, X_i', \dots, X_n) - f(x)| dx \right| \\
&\le \left| \int \widehat{f}_n(x, X_1, \dots, X_n) - \int \widehat{f}_n(x, X_1, \dots, X_i', \dots, X_n) dx \right| \\
&\le \frac{1}{\lfloor \frac{n}{N} \rfloor} \int |f_N(x, X_{(k-1)N+1}, \dots, X_{kN}) - f_N(x, X_{(k-1)N+1}, \dots, X_i', \dots, X_{kN})| dx \\
&\quad \text{(if the index } i \text{ is in the } k\text{-th block)} \\
&\le \frac{2}{\lfloor \frac{n}{N} \rfloor},
\end{aligned}$$

where the last inequality follows from the fact that we may assume that the estimate f_N is always a density (i.e., nonnegative and integrates to one, since otherwise with standard operations one may always construct such an estimate by not increasing the L_1 error, see Devroye (1987)), and that the L_1 distance between any two densities is at most 2. Thus, McDiarmid's inequality implies that for any $t > 0$,

$$\mathbf{P}\left\{ \int |\widehat{f}_n - f| - \mathbf{E} \int |\widehat{f}_n - f| > t \right\} \le e^{-t^2/2n(1/\lfloor \frac{n}{N} \rfloor)^2} \le e^{-t^2 n^{1/3}/8}.$$

if $n \ge 3$. Thus, for example, by taking $t = n^{-1/12}$ and using the simplified notation $a_n' = a_{\lfloor n^{1/3} \rfloor} + n^{-1/12}$, for any $f \in \mathcal{F}$,

$$\mathbf{P}\left\{ \int |\widehat{f}_n - f| > a_n' \right\} \le e^{-n^{1/6}/8}.$$

The second step is to extend $\widehat{f}_n$ so that it becomes universally consistent. Without loss of generality we may assume again that $\widehat{f}_n$ is indeed a density. Let ξ_n be an arbitrary strongly universally consistent density estimate. We proceed as follows: Split the available data into two equal parts $X_1, \dots, X_{n/2}$ and $X_{n/2+1}, \dots, X_n$. Based on the first half construct the estimates $\widehat{f}_{n/2}$ and $h_{n/2}$. Define the "Yatracos set" (see Yatracos (1985) or Devroye and Lugosi (1997))

$$A_{n/2} = \left\{ x : \widehat{f}_{n/2}(x) > \xi_{n/2}(x) \right\},$$

and use the second half of the data to calculate the empirical probability

$$\mu_{n/2}(A_{n/2}) = \frac{2}{n} \sum_{j=n/2+1}^{n} I_{\{X_j \in A_{n/2}\}}.$$

The density estimate g_n is defined by

$$g_n = \begin{cases} \widehat{f}_{n/2} & \text{if } \left| \mu_{n/2}(A_{n/2}) - \int_{A_{n/2}} \widehat{f}_{n/2} \right| < \left| \mu_{n/2}(A_{n/2}) - \int_{A_{n/2}} \xi_{n/2} \right|; \\ \xi_n & \text{otherwise.} \end{cases}$$

We need to show that g_n is universally consistent, and within $\mathcal{F}$ is has a uniform rate of convergence. We start with the case $f \in \mathcal{F}$. Define $c_n = 2a'_{n/2} + 4n^{-1/3}$. If $\int |\widehat{f}_{n/2} - \xi_{n/2}| \le c_n$ then by the triangle inequality

$$\int |g_n - f| \le \int |\widehat{f}_{n/2} - f| + c_n.$$

Otherwise, if $\int |\widehat{f}_{n/2} - \xi_{n/2}| = 2\left(\int_{A_{n/2}} \widehat{f}_{n/2} - \int_{A_{n/2}} \xi_{n/2}\right) > c_n$, observing that

$$\left|\int_{A_{n/2}} f - \int_{A_{n/2}} \widehat{f}_{n/2}\right| \le \frac{1}{2}\int |\widehat{f}_{n/2} - f|$$

by Scheffé's theorem, we have $g_n = \widehat{f}_{n/2}$ whenever

$$\left|\mu_{n/2}(A_{n/2}) - \int_{A_{n/2}} f\right| < \frac{c_n}{4} - \frac{1}{2}\int |\widehat{f}_{n/2} - f|.$$

Summarizing the two cases, we see that for all $f \in \mathcal{F}$,

$$\begin{aligned}
\mathrm{P}\left\{\int |g_n - f| > a'_{n/2} + c_n\right\} &\le \mathrm{P}\left\{\int |g_n - f| > \int |\widehat{f}_{n/2} - f| + c_n\right\} + e^{-(n/2)^{1/6}/8} \\
&\le \mathrm{P}\left\{\left|\mu_{n/2}(A_{n/2}) - \int_{A_{n/2}} f\right| > \frac{c_n}{4} - \frac{a'_{n/2}}{2}\right\} + 2e^{-(n/2)^{1/6}/8} \\
&\quad \text{(by the above argument)} \\
&= \mathrm{P}\left\{\left|\mu_{n/2}(A_{n/2}) - \int_{A_{n/2}} f\right| > n^{-1/3}\right\} + 2e^{-(n/2)^{1/6}/8} \\
&\le e^{-n^{1/3}} + 2e^{-(n/2)^{1/6}/8} \\
&\quad \text{(by Hoeffding's inequality, 1963).}
\end{aligned}$$

Therefore, taking $b_n = a'_{n/2} + c_n$, we have that, indeed, for every $f \in \mathcal{F}$,

$$\sum_{n=1}^{\infty} \mathrm{P}\left\{\int |g_n - f| > b_n\right\} < \infty.$$

Now it remains to show that g_n is strongly universally consistent. This may be done in a similar way: If $\int |\widehat{f}_{n/2} - \xi_{n/2}| \le 2\left(\int |\xi_{n/2} - f| + n^{-1/3}\right)$ then by the triangle inequality,

$$\int |g_n - f| \le 3\int |\xi_{n/2} - f| + 2n^{-1/3}.$$

Otherwise, by Scheffé's theorem, $\int_{A_{n/2}} \widehat{f}_{n/2} - \int_{A_{n/2}} \xi_{n/2} > \int |\xi_{n/2} - f| + n^{-1/3}$ and

$$\left|\int_{A_{n/2}} f - \int_{A_{n/2}} \xi_{n/2}\right| \le \frac{1}{2}\int |\xi_{n/2} - f|.$$

Therefore, $g_n = \xi_{n/2}$ whenever

$$\left|\mu_{n/2}(A_{n/2}) - \int_{A_{n/2}} f\right| < \frac{1}{2} n^{-1/3}.$$

Therefore, in all cases, since the L_1 distance is bounded by 2,

$$\int |g_n - f| \le 3\int |\xi_{n/2} - f| + 2n^{-1/3} + 2I_{\left\{\left|\mu_{n/2}(A_{n/2}) - \int_{A_{n/2}} f\right| \ge (1/2)n^{-1/3}\right\}}.$$

By the strong universal consistency of ξ_n and Hoeffding's inequality, all terms on the right-hand side converge to zero almost surely. This concludes the proof of Lemma 5 and Theorem 2. $\square$

§10. Non-discernible classes

In this section we establish sufficient conditions for the nondiscernibility of a class of densities, and show several examples of nondiscernible classes.

We first generalize the definition of discernibility: if $\mathcal{X}$ is a class of densities, then we say that $\mathcal{F}$ is **discernible with respect to** $\mathcal{X}$ if there exists a consistent classification rule to decide whether $f \in \mathcal{F} \cap \mathcal{X}$ or $f \in \mathcal{F}^c \cap \mathcal{X}$. Clearly, if $\mathcal{F}$ is not discernible with respect to $\mathcal{X}$, then $\mathcal{F}$ is not discernible (with respect to the class of all densities).

It will be convenient to work, instead of densities, with the inverse of their corresponding cumulative distribution function: Any density f is uniquely determined by a monotonically increasing function $G : (0,1) \to (-\infty, \infty)$ defined by $G = F^{-1}$, where $F(x) = \int_{-\infty}^{x} f(z)dz$. If $U_1, U_2, \ldots$ is a sequence of independent uniform $[0,1]$ random variables, then $G(U_1), G(U_2), \ldots$ is a sequence of i.i.d. random variables with density f. This is the coupling between samples from different distributions which will be used in the proof below.

THEOREM 8. *Let $\mathcal{F}, \mathcal{X}$ be classes of densities. Denote the class of all inverse distribution functions corresponding to densities in $\mathcal{X}$ by $\mathcal{A}$, and to those in $\mathcal{F} \cap \mathcal{X}$ by $\mathcal{G}$. Assume that there exist two sets of functions $\mathcal{B} \subset \mathcal{G}$ and $\mathcal{C} \subset \mathcal{A} - \mathcal{G}$ with the following property:*

(1) there exists a family of subsets S_ϵ of $(0,1)$ indexed by real numbers $\epsilon \in (0,1)$ such that $\lambda(S_\epsilon) \leq \epsilon$ (λ denotes the Lebesgue measure);

(2) if $\epsilon_1 > \epsilon_2$ then $S_{\epsilon_1} \supseteq S_{\epsilon_2}$;

(3) for any $\epsilon \in (0,1)$ and $G \in \mathcal{B}$ there exists a $H \in \mathcal{C}$ such that $G(x) = H(x)$ for all $x \notin S_\epsilon$.

(4) for any $\epsilon \in (0,1)$ and $H \in \mathcal{C}$ there exists a $G \in \mathcal{B}$ such that $G(x) = H(x)$ for all $x \notin S_\epsilon$.

Then $\mathcal{F}$ is not discernible with respect to $\mathcal{X}$.

PROOF. Let $U_1, U_2, \ldots$ be an i.i.d. sequence of uniform $[0,1]$ random variables, from which we obtain all samples for all distributions by the inverse distribution function transformation. We may represent this sequence by the probability element ω. Assume that there exists a consistent classification rule T_n. Then for any density $f \in \mathcal{F} \cap \mathcal{X}$, and almost all ω (i.e., with probability one) there exists an integer $N(\omega)$ such that

$$T_n(X_1, \ldots, X_n) = 1 \quad \text{if } n > N(\omega)$$

and for any density $f \in \mathcal{F}^c \cap \mathcal{X}$, and almost all ω there exists an integer $N(\omega)$ such that

$$T_n(X_1, \ldots, X_n) = 0 \quad \text{if } n > N(\omega).$$

We will construct a density such that, with probability more than 1/2, $T_n(X_1, \ldots, X_n) = 0$ for infinitely many n and $T_n(X_1, \ldots, X_n) = 1$ for infinitely many n, which is a contradiction.

We use the coupling defined in the introduction of this section. Let $\delta_k = 2^{-k-2}$, $k = 1, 2, \ldots$, and let $G_1 \in \mathcal{B}$ be arbitrary. Then there exists an integer N_1 such that

$$\mathbf{P}\left\{T_n(G_1(U_1), \ldots, G_1(U_n)) = 1 \text{ for all } n \geq N_1\right\} > 1 - \delta_1$$

(see, e.g., Royden, 1968, p. 70, Problem 23.a). Choose $\epsilon_1 > 0$ such that $(1-\epsilon_1)^{N_1} > 1-\delta_1$, and consider a function $G_2 \in \mathcal{C}$ which agrees with G_1 on $(0,1) - S_{\epsilon_1}$. Then

$$\mathrm{P}\left\{G_1(U_1) = G_2(U_1), \ldots, G_1(U_{N_1}) = G_2(U_{N_1})\right\} > 1 - \delta_1,$$

and

$$\mathrm{P}\left\{T_n(G_2(U_1), \ldots, G_2(U_{N_1})) = 1\right\} > 1 - 2\delta_1.$$

Now similarly, since $G_2 \in \mathcal{C}$, there exists an integer $N_2 > N_1$ such that

$$\mathrm{P}\left\{T_n(G_2(U_1), \ldots, G_2(U_n)) = 0 \text{ for all } n \geq N_2\right\} > 1 - \delta_2.$$

Next we choose $\epsilon_2 > 0$ such that $(1-\epsilon_2)^{N_2} > 1-\delta_2$, and consider a function $G_3 \in \mathcal{B}$ which agrees with G_2 on $(0,1) - S_{\epsilon_2}$. We continue this procedure such that $G_k \in \mathcal{B}$ for odd k and $G_k \in \mathcal{C}$ for even k, and G_k agrees with G_{k-1} on $(0,1) - S_{\epsilon_{k-1}}$, where $(1-\epsilon_{k-1})^{N_{k-1}} > 1-\delta_{k-1}$, and N_k is chosen such that

$$\mathrm{P}\left\{T_n(G_k(U_1), \ldots, G_k(U_n)) = I_{\{k \text{ is odd}\}} \text{ for all } n \geq N_k\right\} > 1 - \delta_k.$$

Then it follows from assumption (2) that the sequence of these functions $G_1, G_2, \ldots$ converges pointwise to some function $G \in \mathcal{A}$. Also,

$$\mathrm{P}\left\{T_{N_k}(G(U_1), \ldots, G(U_{N_k})) = I_{\{k \text{ is odd}\}} \text{ for all } k = 1, 2, \ldots\right\} > 1 - \sum_{k=1}^{\infty} 2\delta_k = \frac{1}{2},$$

and the proof is finished. □

Example 18: Boundedness of the support of a density is not discernible. Let $\mathcal{F}$ be the class of all densities whose support is bounded. Then Theorem 8 implies that $\mathcal{F}$ is not discernible. To see this, simply take $S_\epsilon = (1-\epsilon, 1)$, and let $\mathcal{B}$ be the family of inverse distribution functions corresponding to positive bounded random variables with a density, and let $\mathcal{C}$ be the family corresponding to positive unbounded random variables with a density. Then clearly, functions in $\mathcal{B}$ are bounded and functions in $\mathcal{C}$ are unbounded, and for any bounded function $G \in \mathcal{B}$ and $\epsilon \in (0,1)$ there exists an unbounded $H \in \mathcal{C}$ such that G and H agree on $(0, 1-\epsilon)$ and vice versa, and therefore the condition of Theorem 8 is satisfied. It is also easy to see that one cannot construct consistent classification rules for boundedness of the support with respect to the following classes: all continuous densities, all Lipschitz densities, and all unimodal densities. In particular, unimodality is discernible, but not bounded support once it is known that a density is unimodal.

Example 19: Boundedness of a density is not discernible. Let $\mathcal{F}$ be the class of all densities with $\operatorname{ess\,sup} f < \infty$. Then $\mathcal{F}$ is not discernible. We show this, via Theorem 8, by proving that $\mathcal{F}$ is not discernible with respect to $\mathcal{X}$, the class of all monotonically decreasing densities on $[0,\infty)$ that are continuous on $(0,\infty)$. Then all inverse distribution functions $G \in \mathcal{A}$ corresponding to densities in $\mathcal{X}$ are concave increasing functions with $G(0) = 0$, $G'(0) \geq 0$, and $\liminf_{t\downarrow 0} G'(t) = \infty$ if and only if the density is unbounded. Take $S_\epsilon = (0, \epsilon)$, and continue a function G for an unbounded density on $[0, \epsilon]$ by a parabola through $(0,0)$ and $(\epsilon, G(\epsilon))$ with derivative $G'(\epsilon)$ at ϵ (which corresponds to a G for a bounded continuous and monotonically decreasing density on $[0,\infty)$). Similarly, continue a function G for a bounded density on $[0,\epsilon)$ by a quadratic Bezier spline (Farin, 1993) having derivative ∞ at 0, $G'(\epsilon)$ at ϵ, and with control point at $(0, a)$, where a is the place where the line through $(\epsilon, G(\epsilon))$ with derivative $G'(\epsilon)$ crosses the y-axis. It is clear that all four conditions of Theorem 8 are satisfied, and therefore $\mathcal{F}$ is not discernible with respect to $\mathcal{X}$.

Example 20: Square integrability of a density is not discernible. Consider the class $\mathcal{F}$ of all densities f for which $\int f^2 dx < \infty$. Then the same argument as in the previous example shows that $\mathcal{F}$ is not discernible with respect to $\mathcal{X}$, the class of monotonically decreasing densities on $[0,\infty)$ that are continuous on $(0,\infty)$ (and thus have possibly an infinite peak at the origin). For the construction involving continuations, just note that a density with $\int f^2 = \infty$ is continued on $[0,\epsilon)$ as in Example 19 by a bounded density (which necessarily has $\int f^2 < \infty$). The other continuation argument is trickier and requires a density whose behavior near the origin is as $c/2\sqrt{x}$ (and thus is not square integrable), and whose distribution function G evolves as $c\sqrt{x}$. By choice of c, such a continuation exists that meets the requirements of continuity and monotonicity as well.

EXAMPLE 21: CONTINUITY OF A DENSITY IS NOT DISCERNIBLE. Let $\mathcal{F}$ be the class of densities such that each $f \in \mathcal{F}$ has a version which is continuous at 0. This class is not discernible. To see this, we apply Theorem 8 in the following setup: $\mathcal{X}$ is the class of all strictly monotonically increasing densities on $[0,1]$, continuous on $(0,\infty)$, with $f(1) = 1$ and $f(x) = (2-x)_+, x > 1$. Taking again $S_\epsilon = (0,\epsilon)$, it is easy to see that conditions (3) and (4) of Theorem 8 are satisfied, which implies the non-discernibility of $\mathcal{F}$ with respect to $\mathcal{X}$. Indeed, we apply the same extension arguments from the previous examples, and note that the distribution functions G should be convex and increasing on $[0,1]$, with continuous derivatives there, and with $G'(0) = 0$ if and only if f is continuous at 0 (and $f(0) = 0$). Thus, within the class, we may always continue a G with $G'(0) > 0$ on $[0,\epsilon]$ by one having $G'(0) = 0$ using a quadratic Bezier spline having the specified derivatives at 0 and ϵ. Similarly, a function G with $G'(0) = 0$ may be continued on $[0,\epsilon)$ by a quadratic Bezier spline B having $B'(0) = G(\epsilon)/(2\epsilon) > 0$ and $B'(\epsilon) = G'(\epsilon)$. It should be easy to generalize and conclude that the class of densities with $\int f^r < \infty$ (for fixed $r > 1$) is not discernible with respect to the decreasing densities on $[0,1]$. The same goes for conditions like $\int f \log^a(1+f) < \infty$ for any $a > 0$, and indeed many other functionals.

EXAMPLE 22: LIPSCHITZ CONTINUITY OF A DENSITY IS NOT DISCERNIBLE. Let $\mathcal{F}$ be the class of densities such that each $f \in \mathcal{F}$ has a version which is Lipschitz at 0. This class is not discernible. To see this, we consider $\mathcal{X}$ as in the previous example and ask additionally for continuity on the whole line (so that for each density, $f(0) = 0$). We need to slightly modify the continuation argument. If G corresponds to a density in $\mathcal{F} \cap \mathcal{X}$, then, near the origin, $G(x) \le Cx^2$ for some constant C. Such a G may be continued on $[0,\epsilon)$ by another G behaving as $cx^{3/2}$ near the origin, where $c > 0$ is picked appropriately, while maintaining the other restrictions imposed by $\mathcal{X}$. Note that the corresponding density behaves as $\sqrt{x}$ near the origin, and is thus not Lipschitz. If G corresponds to a density in $\mathcal{F}^c \cap \mathcal{X}$, then, near the origin, we may continue G by another function within $\mathcal{X}$ that behaves as Cx^2 near the origin for an appropriately small but positive C. Thus, within $\mathcal{X}$, Lipschitz continuity at even one point is not discernible.

We leave it to the reader as a simple exercise now to verify that we cannot test whether $\mathbf{E}|X|^a < \infty$ for fixed $a > 0$ with respect to the class of bounded symmetric unimodal densities (by continuations of the tails by finite support or heavy-tailed pieces). Similarly, the finiteness of a moment generating function at a given point cannot be tested. It is also easy to see that the L_1 ball $\mathcal{F}_\epsilon = \{f : \int |f - f_0| \le \epsilon\}$ centered at a fixed density f_0 is non-discernible.

§11. Non-discernibility of mixture classes

Consider the class $\mathcal{X}$ of densities which may be obtained as densities of random variables of the form $X = F(U,Z)$, where $F : [0,1] \times \{1,2,\ldots\} \to \mathcal{R}$ is a fixed and known measurable function, U is a uniform random variable on $[0,1]$, and Z is an arbitrary positive-integer-valued random variable, independent of U. We assume that F is such that $F(U,Z)$ has a density for all possible Z, and that for any two different distributions of Z, $F(U,Z)$ has different densities. Let $\mathcal{F}$ be the subclass of $\mathcal{X}$ containing all densities such that Z has a finite support.

THEOREM 9. *The class $\mathcal{F}$ defined above is not discernible with respect to $\mathcal{X}$.*

PROOF. We proceed similarly as in Theorem 8. Just like there, we need a coupling between samples drawn from different densities of $\mathcal{X}$. A simple way of doing it is as follows: Let $U_1, V_1, U_2, V_2, \ldots$ be i.i.d. uniform random variables on $[0, 1]$. Given a distribution $(p_1, p_2, \ldots)$ on the set of positive integers, let $Z_i = j$ if and only if $V_i \in \left[\sum_{k=0}^{j-1} p_k, \sum_{k=0}^{j} p_k\right)$. The sample drawn from a density $f \in \mathcal{X}$ is now $X_1, X_2, \ldots$, where $X_i = F(U_i, Z_i)$. The densities in $\mathcal{F}$ are characterized by the distributions $(p_1, p_2, \ldots)$ which have finite support (i.e., only finitely many p_i's are nonzero). Call this class $\mathcal{P}$. Similarly, the class $\mathcal{X} - \mathcal{F}$ corresponds to the class $\mathcal{Q}$ of distributions with infinite support.

Assume that the statement is false, and there exists a classification rule T_n such that for any $f \in \mathcal{F}$, with probability one, there exists an integer $N(\omega)$ such that

$$T_n(X_1, \ldots, X_n) = 1 \quad \text{if } n > N(\omega),$$

and for any $f \in \mathcal{X} - \mathcal{F}$, with probability one, there exists an integer $N(\omega)$ such that

$$T_n(X_1, \ldots, X_n) = 0 \quad \text{if } n > N(\omega).$$

Let $\delta_k = 2^{-k-2}$, $k = 1, 2, \ldots$.

Let $(p_1^{(1)}, p_2^{(1)}, \ldots) \in \mathcal{P}$ be arbitrary. Then there exists an integer N_1 such that

$$\mathbf{P}\left\{T_n(X_1^{(1)}, \ldots, X_n^{(1)}) = 1 \text{ for all } n \geq N_1\right\} > 1 - \delta_1,$$

where $X_1^{(1)}, \ldots, X_n^{(1)}$ is the sample drawn from the corresponding density as described above. Next consider another distribution $(p_1^{(2)}, p_2^{(2)}, \ldots) \in \mathcal{Q}$ such that

$$\mathbf{P}\left\{X_1^{(1)} = X_1^{(2)}, \ldots, X_n^{(1)} = X_n^{(2)}\right\} > 1 - \delta_1.$$

Such a distribution may be constructed by choosing $\epsilon_1 > 0$ such that $(1 - \epsilon_1)^{N_1} > 1 - \delta_1$, and defining $p_i^{(2)} = p_i^{(1)}$ for all $i < m$, where m is the largest integer such that $p_m^{(2)} > 0$, $p_m^{(2)} = \min(0, p_m^{(1)} - \epsilon_1)$, and $p_i^{(2)} = 2^{-(i-m+1)} \min(\epsilon_1, p_m^{(1)})$ for all $i \geq m$.

This step may be iterated just like in Theorem 8, leading to the construction of a limiting density in $\mathcal{X} - \mathcal{F}$ for which, with probability greater than 1/2, the classification rule T_n fails for every odd n. The proof is finished. □

EXAMPLE 23: NORMAL MIXTURES. It is not discernible whether a normal mixture density has finitely or infinitely many components. This may be straightforwardly cast in the framework of Theorem 9: let $\mathcal{X}$ be the class of all densities of the form

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{\infty} \frac{p_i}{\sigma_i} e^{-\frac{(x-m_i)^2}{2\sigma_i^2}},$$

where $p_1, p_2, \ldots$ is a probability vector, $\sigma_1, \sigma_2, \ldots$ are positive numbers, and $m_1, m_2, \ldots$ are arbitrary real numbers. Assume that all parameters except the p_i's are known, an that no two pairs (m_i, σ_i) are identical. This class is $\mathcal{X}$. According to Theorem 9, the class $\mathcal{F}$ consisting of all densities of the above form such that $p_i > 0$ for only finitely many i's is not discernible with respect to $\mathcal{X}$. From this, it follows certainly that the class of all finite mixtures of normal densities is not discernible (here, the parameters m_i and σ_i are unrestricted).

EXAMPLE 24: CHARACTERISTIC FUNCTIONS WITH COMPACT SUPPORT. Let $\mathcal{F}$ be the class of all densities whose characteristic function φ has bounded support. This class is not discernible. To see this, we take densities f_n with characteristic function $(1 - |t|/n)_+$, and let $\mathcal{X}$ be the class of all mixtures of finitely or infinitely many components f_n. Then the class of finite mixtures of f_n's is not discernible with respect to $\mathcal{X}$, by Theorem 9.

EXAMPLE 25: CHARACTERISTIC FUNCTIONS WITH EXPONENTIALLY DECREASING TAILS. Let $\mathcal{F}$ be the class of all densities whose characteristic function φ drops off exponentially quickly, i.e., for which $|\varphi(t)| \le Ce^{-a|t|}$ for positive constants C, a. This class is not discernible. To see this, we take Cauchy densities f_n with characteristic function $e^{-|t|/n}$, and let $\mathcal{X}$ be the class of all mixtures of finitely or infinitely many components f_n. Then the class of finite mixtures of f_n's is not discernible with respect to $\mathcal{X}$, by Theorem 9. However, for any inifinite mixture, the mixture density does not have an exponentially decaying characteristic function.

§12. References

T. M. Cover, "On determining the irrationality of the mean of a random variable," *Annals of Statistics*, vol. 1, p. 862,, 1973.

A. Dembo and Y. Peres, "A topological criterion for hypothesis testing," *Annals of Statistics*, vol. 22, pp. 106–117, 1994.

L. Devroye, "On arbitrarily slow rates of global convergence in density estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 62, pp. 475–483, 1983.

L. Devroye, "A simple algorithm for generating random variates with a log-concave density," *Computing*, vol. 33, pp. 247–257, 1984.

L. Devroye, "A note on the L1 consistency of variable kernel estimates," *Annals of Statistics*, vol. 13, pp. 1041–1049, 1985.

L. Devroye, *A Course In Density Estimation*, Birkhäuser Verlag, Boston, 1987.

L. Devroye, "Asymptotic performance bounds for the kernel estimate," *Annals of Statistics*, vol. 16, pp. 1162–1179, 1988.

L. Devroye, "Nonparametric density estimates with improved performance on given sets of densities," *Statistics (Mathematische Operationsforschung und Statistik)*, vol. 20, pp. 357–376, 1989.

L. Devroye, "Universal smoothing factor selection in density estimation: theory and practice (with discussion)," *Test*, vol. 6, pp. 223–320, 1997.

L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

L. Devroye and G. Lugosi, "A universally acceptable smoothing factor for kernel density estimation," *Annals of Statistics*, vol. 24, pp. 2499–2512, 1996.

L. Devroye and G. Lugosi, "Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes," *Annals of Statistics*, vol. 25, pp. 2626–2637, 1997.

L. Devroye, G. Lugosi, and F. Udina, "Inequalities for a new data-based method for selecting nonparametric densitys," in M.L. Puri (editor), *Festschrift in Honour of George Roussas*, VSP International Science Publishers, 2000.

D. L. Donoho, "One-sided inference about functionals of a density," *Annals of Statistics*, vol. 16, pp. 1390–1420, 1988.

W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

W. Hoeffding and J. Wolfowitz, "Distinguishabiloity of sets of distributions," *Annals of Mathemetical Statistics*, vol. 29, pp. 700–718, 1958.

J. Koplowitz, "Abstracts of papers," in: *International Symposium on Information Theory*, p. 64, Cornell University, Ithaca, NY, 1977.

S. R. Kulkarni and O. Zeitouni, "A general classification rule for probability measures," *Annals of Statistics*, vol. 23, pp. 1393–1407, 1995.

L. LeCam and L. Schwartz, "A necessary and sufficient condition for the existence of consistent estimates," *Annals of Mathemetical Statistics*, vol. 31, pp. 140–150, 1960.

G. G. Lorentz, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *Annals of Probability*, vol. 18, pp. 1269–1283, 1990.

C. McDiarmid, "On the method of bounded differences," in: *Surveys in Combinatorics*, ed. J. Siemons, vol. 141, pp. 148–188, London Mathematical Society Lecture Note Series, Cambridge University Press, 1989.

H. L. Royden, *Real Analysis*, Macmillan Publishing Co., New York, 1968.

V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, pp. 264–280, 1971.

R. L. Wheeden and A. Zygmund, *Measure and Integral*, Marcel Dekker, New York, 1977.

Y. G. Yatracos, "Rates of convergence of minimum distance estimators and Kolmogorov's entropy," *Annals of Statistics*, vol. 13, pp. 768–774, 1985.

Y. G. Yatracos, "A note on L_1 consistent estimation," *Canadian Journal of Statistics*, vol. 16, pp. 283–292, 1988.