Quickest Detection of Abrupt Changes for a Class of Random Processes

George V. Moustakides, *Senior Member, IEEE*

***Abstract*— We consider the problem of quickest detection of abrupt changes for processes that are not necessarily independent and identically distributed (i.i.d.) before and after the change. By making a very simple observation that applies to most well-known optimum stopping times developed for this problem (in particular CUSUM and Shiryayev–Roberts stopping rule) we show that their optimality can be easily extended to more general processes than the usual i.i.d. case.**

***Index Terms*—CUSUM, disruption problem, optimal stopping, quickest detection.**

I. Introduction

Recently, there has been rapidly increasing interest in the area of change detection. This is due to the large number of applications that can be mathematically formulated through this problem as well as the development of theoretically and computationally tractable techniques that can provide the corresponding solutions. Such applications include: quality control, system monitoring, vibration monitoring, segmentation of signals, change point problems in economic, medical, seismic and astrophysical time series, etc. Characteristic references and review articles for each application area can be found in [1], [2], and a very complete presentation of the existing methodology in [1].

Despite the abundance of techniques addressing the change detection problem, optimum schemes can be found only for the case where the data are independent and identically distributed (i.i.d.) and the distributions are completely known before and after the change. For dependent data and/or cases where the distributions are partially known, existing methods are either *ad hoc* or, at best, asymptotically optimum. The only exception is the case of finite-state Markov chains, where there exists a nonasymptotic optimality result for one of the possible formulations of the change detection problem [13].

In this correspondence, by making a simple observation that applies to most well-known optimum detection schemes, we are going to extend their optimality property to a class of processes that is not necessarily i.i.d. As will be seen, this class contains a number of interesting examples of processes that are characterized by common dependency structures used to model dependent data. This generalization will, in fact, be possible at almost no cost as far as mathematical analysis is concerned.

Before rigorously presenting the change detection problem and the existing optimum schemes let us first introduce the data model we intend to work with. Assume that we are given a sequence of random variables $\xi_1, \xi_2, \cdots$ and denote by $\{\mathcal{F}_n\}$ the corresponding filtration generated by $\{\xi_n\}$. Consider also that we are given two sequences of conditional probability measures $\{\mathbb{P}_n(\xi_n|\mathcal{F}_{n-1})\}, \{\mathbb{Q}_n(\xi_n|\mathcal{F}_{n-1})\}$ that can describe the statistics of $\{\xi_n\}$ with $\mathbb{Q}_n(\xi_n|\mathcal{F}_{n-1})$ absolutely continuous with respect to $\mathbb{P}_n(\xi_n|\mathcal{F}_{n-1})$ for every $n \geq 1$. If

$$l_n = \frac{d\mathbb{Q}_n(\xi_n|\mathcal{F}_{n-1})}{d\mathbb{P}_n(\xi_n|\mathcal{F}_{n-1})} \tag{1}$$

denotes the Radon–Nikodym derivative of the two conditional measures at time n, we are interested in processes that satisfy the following key condition:

$$\mathbb{P}_n\{\xi_n\colon l_n \leq x|\mathcal{F}_{n-1}\} = F_0(x) \tag{2}$$

for every $x \geq 0$ and every $n \geq 1$. Albeit the process $\{\xi_n\}$ can be nonstationary and dependent, meaning that the conditional probability of ξ_n given $\mathcal{F}_{n-1}$ depends in general on $\mathcal{F}_{n-1}$, with condition (2) we require the conditional probability of l_n to be independent of $\mathcal{F}_{n-1}$ and stationary. The following lemma will, in fact, guarantee that the corresponding process $\{l_n\}$ is i.i.d. under both measures induced by the corresponding sequences $\{\mathbb{P}_n\}, \{\mathbb{Q}_n\}$.

Lemma 1: Let $\{l_n\}$ be the sequence of random variables defined by (1); then this process is i.i.d. under both probability measures induced by the two sequences of conditional probability measure $\{\mathbb{P}_n\}$ and $\{\mathbb{Q}_n\}$ with marginal distribution functions equal to $F_0(x)$ and $F_1(x) = \int_0^x z\, dF_0(z)$, respectively.

Proof: Let us first compute the multivariate distribution of the random variables $l_1, \cdots, l_n$, under the probability measure induced by the conditional measures $\{\mathbb{P}_n\}$. Using (2) and the fact that the event $\{l_n \leq x\}$ is $\mathcal{F}_n$ measurable we can write

$$\begin{aligned} \Pr\{l_n \leq x_n, \cdots, l_1 \leq x_1\} &= \Pr\{l_n \leq x_n|\mathcal{F}_{n-1}\} \cdots \Pr\{l_1 \leq x_1|\mathcal{F}_0\} \\ &= \mathbb{P}_n\{\xi_n\colon l_n \leq x_n|\mathcal{F}_{n-1}\} \cdots \mathbb{P}_1\{\xi_1\colon l_1 \leq x_1|\mathcal{F}_0\} \\ &= F_0(x_n) \cdots F_0(x_1) \end{aligned} \tag{3}$$

which proves that $\{l_n\}$ is i.i.d. with corresponding marginal distribution equal to $F_0(x)$. To prove that this is also the case for the sequence $\{\mathbb{Q}_n\}$, it is enough to show that a similar relation as in (2) holds under this alternative sequence of conditional probability measures, specifically that we can have

$$F_1(x) = \mathbb{Q}_n\{\xi_n\colon l_n \leq x|\mathcal{F}_{n-1}\} \tag{4}$$

By a simple application of change of measures and using (2) it is easy to show that indeed the above equation is valid with $F_1(x) = \int_0^x z\, dF_0(z)$. □

The requirement imposed by condition (2) is, of course, very restrictive; however, as we will see in Section III, there are several interesting processes that can satisfy it. On the other hand, condition (2) turns out to be sufficient for proving optimality for most well-known stopping times encountered in the literature (for example, CUSUM and the Shiryayev–Roberts stopping rule). This is because these tests use only the process $\{l_n\}$ to form their test statistics and the corresponding proofs rely only on the fact that this process is i.i.d. Since for processes $\{\xi_n\}$ satisfying (2), the process $\{l_n\}$ is comprised of mutually independent random variables, the optimality property of the corresponding stopping times is clearly preserved.

II. Optimal Stopping Times

In this section we are going to describe the change detection problem (also known as disruption problem) and briefly present the stopping times that are known to optimally solve it.

Manuscript received August 19, 1997; revised March 27, 1998. This work was completed while the author was a Visiting Fellow at the Department of Electrical Engineering, Princeton University, Princeton, NJ 08854.

The author is with the Department of Computer Engineering and Informatics, University of Patras, 26900 Patras, Greece.

Publisher Item Identifier S 0018-9448(98)05124-4.

We assume that we are given a sequence of random variables $\xi_1, \xi_2, \cdots$ and for some unknown time $m \geq 1$ the random variables $\xi_1, \cdots, \xi_{m-1}$ are distributed according to the conditional probability measures $\mathbb{P}_1, \cdots, \mathbb{P}_{m-1}$ whereas the random variables $\xi_m, \xi_{m+1}, \cdots$ according to $\mathbb{Q}_m, \mathbb{Q}_{m+1}, \cdots$. We are interested in detecting the time of change m. Detection is signaled through a stopping time N that attempts to minimize the detection delay controlling at the same time the false alarm.

Let $\mathbb{E}_m$ denote expectation under the probability measure obtained when there is a change at time m and $\mathbb{E}_\infty$ expectation under no change condition $(m = \infty)$. Let also all stopping times, we like to consider, be adapted to the filtration $\{\mathcal{F}_n\}$ generated by the sequence $\{\xi_n\}$.

There exist several formulations in the literature regarding the change detection problem. A Bayesian approach proposed by Shiryayev in [10, pp. 193–198] assumes a geometric prior on the change time m of the form $\Pr\{m = 0\} = \pi$ and

$$\Pr\{m = n\} = (1-\pi)(1-p)^{n-1}p, \qquad n \geq 1.$$

The optimum stopping rule is required to minimize the risk function $J(N) = \mathbb{E}\{\mathbb{1}_{\{N < m\}} + c(N-m)^+\}$. Shiryayev proved that, for the case where $\{\xi_n\}$ is i.i.d. before and after the change, the optimum stopping rule consists in stopping the first time the posterior probability π_n (that the change has occurred) exceeds some constant threshold. Since π_n satisfies the recursion

$$\pi_n = \frac{\pi_{n-1}l_n + (1-\pi_{n-1})pl_n}{\pi_{n-1}l_n + (1-\pi_{n-1})pl_n + (1-\pi_{n-1})(1-p)} \tag{5}$$

we can see that its computation depends only on the process $\{l_n\}$; therefore, the proof of optimality can be also valid for processes satisfying condition (2).

A non-Bayesian formulation suggested by Lorden [3] consists in finding the stopping time that minimizes

$$J(N) = \sup_{m \geq 1} \text{ ess sup } \mathbb{E}\{(N-m+1)^+ | \mathcal{F}_{m-1}\}$$

among all stopping times that satisfy the false alarm constraint $\mathbb{E}_\infty\{N\} \geq \gamma$ for some given $\gamma > 0$. Lorden showed in [3] that the CUSUM test, introduced by Page [5], is asymptotically optimum. Nonasymptotic optimality of the CUSUM test was first shown in [4] and a different proof based on a Bayesian saddle point formulation was provided in [8]. The CUSUM test is defined through the stopping time $N_C = \inf_n\{n: T_n \geq \mu\}$ where the statistic T_n satisfies the recursion

$$T_n = \max\{T_{n-1}, 1\}l_n \tag{6}$$

and μ is a threshold selected to meet the false alarm constraint with equality. All three proofs consider the case where $\{\xi_n\}$ is i.i.d. before and after the change but again we see that the test statistic T_n depends only on $\{l_n\}$; therefore, the proofs can be easily extended to cover processes satisfying (2).

An alternative non-Bayesian formulation was proposed by Pollak and Seigmund in [7] consisting in minimizing the functional

$$J(N) = \sup_{m \geq 1} \mathbb{E}_m\{N - m | N \geq m\}$$

under the false alarm constraint $\mathbb{E}_\infty\{N\} \geq \gamma$. Pollak in [6] proved that the Shiryayev–Roberts stopping rule [9] asymptotically minimizes $J(N)$ when the sequence $\{\xi_n\}$ is i.i.d. before and after the change. Nonasymptotic optimality of a modified version of the Shiryayev–Roberts stopping rule was recently presented by Yakir [14] for the i.i.d. case. The corresponding stopping time is defined as $N_S = \inf_n\{n: S_n \geq \mu\}$ where the statistics S_n is defined through the recursion

$$S_n = (1 + S_{n-1})l_n \tag{7}$$

and μ is a threshold selected to meet the false alarm constraint with equality. Again, the proof can be shown to be valid for processes satisfying (2).

Remark. If instead of the change detection problem we consider the sequential hypotheses problem then, for exactly the same reasons, we can extend the optimality property of the Sequential Probability Ratio Test (SPRT) [11], [12] to processes satisfying (2).

The almost obvious generalization of optimality of most well-known stopping times to processes satisfying (2) has, surprisingly, a number of interesting applications that cannot be easily seen to reduce to the usual i.i.d. case. We present several such examples in the next section.

III. Examples

In this section we are going to present combinations of conditional probability measures for which key condition (2) is satisfied and therefore these cases reduce to the usual i.i.d. case.

A. Markov Chains

As was stated in the Introduction, this is the only dependency model for which there exists a nonasymptotic optimum detection scheme. Specifically, in [13] the change detection problem is solved under the Bayesian formulation of geometric priors on the change time. The optimum scheme is similar to the original one introduced by Shiryayev, only now the threshold is state-dependent.

Here, in order to satisfy key condition (2), we are going to introduce Markov chains with a special dependency structure. Specifically, consider a Markov chain that has the same finite state space X before and after the change and assume, without loss of generality, that $X = \{1, \cdots, K\}$. Let $P = [p_1 \cdots p_K]^t$ and $Q = [q_1 \cdots q_K]^t$ denote the two transition matrices before and after the change with $p_i^t, q_i^t, i = 1, \cdots, K$, their corresponding rows and superscript "t" denoting transpose.[1] If $T_i, i = 1, \cdots, K$, are K permutation matrices (not necessarily distinct) and there exist vectors p and q such that

$$p_i = T_i p, \quad q_i = T_i q, \qquad i = 1, \cdots, K \tag{8}$$

then (2) is satisfied. The above condition implies that corresponding rows of the matrices P and Q are obtained by applying the same permutation on the elements of the vectors p and q, respectively. Notice that if all permutation matrices are equal to the identity matrix then this reduces to the usual i.i.d. case.

To show (2) we can see that l_n can be represented as a matrix L which can be computed by dividing the two matrices Q and P element-wise. For the matrix L we then have that its rows can be obtained by applying the same permutation matrices $T_i, i = 1, \cdots, K$, to a vector l which is equal to the element-wise division of the vectors q and p. Since the same permutations apply to P and L we conclude that l_n can take only upon the K values of the vector l; furthermore, each such value has a P probability of occurrence equal to the corresponding element of the vector p, this being true independently of $\mathcal{F}_{n-1}$, i.e., the row (state) the process was in at time $n-1$.

What was said above continuous to be valid if the vectors p, q are kept constant but the permutation matrices are time-dependent, thus generating a nonhomogeneous Markov chain.

[1] We assume that the ith row contains the transition probabilities for state i.

Let us consider the simple but illustrative special case of a Markov chain with two states. For any a, b in the interval $[0, 1]$, in order for (2) to be valid, the transition matrices P, Q can either be

$$P = \begin{bmatrix} a & 1-a \\ a & 1-a \end{bmatrix} \qquad Q = \begin{bmatrix} b & 1-b \\ b & 1-b \end{bmatrix} \tag{9}$$

corresponding to the usual i.i.d. case, or

$$P = \begin{bmatrix} a & 1-a \\ 1-a & a \end{bmatrix} \qquad Q = \begin{bmatrix} b & 1-b \\ 1-b & b \end{bmatrix}. \tag{10}$$

These are the only two possibilities for a homogeneous chain with two states. In particular, we can see that the interesting symmetric case defined in (10) belongs to the class of processes satisfying key condition (2). For nonhomogeneous chains, the transition matrices become a function of time and can alternate between (9) and (10).

There is a point that needs to be stressed here. Notice that we require condition (2) to be valid for every $n \geq 1$. It is clear that, for this example, we have validity of the key condition for $n \geq 2$. In order for (2) to be true for $n = 1$ as well, the initial probability measures of the chain before and after the change must be, respectively, equal to p and q (or the same permuted version of these two vectors).

B. AR Processes

Let us consider a process $\{\xi_n\}$ that evolves, before and after the change, on some set C_0 that has finite Lebesgue measure. Let us also assume that before the change $\{\xi_n\}$ is i.i.d. and uniformly distributed on C_0 and after the change that it has an AR dependency structure of the form

$$\xi_n = \alpha \xi_{n-1} + w_n, \qquad \xi_0 = 0 \tag{11}$$

with $|\alpha| < 1$ and $\{w_n\}$ i.i.d. with a probability density equal to $f_1(w)$. We must stress that under the aforementioned model the support of $f_1(w)$ must be constraint in order for the process $\{\xi_n\}$ to evolve on C_0. If, for example, $C_0 = [-1, 1]$ then this requirement is fulfilled whenever the support of $f_1(w)$ is a subset of the interval $[-(1-|\alpha|), (1-|\alpha|)]$.

Under the above constraint we can show that condition (2) is true. Indeed, with ν denoting the Lebesgue measure and by making use of its translation invariance property we have

$$\begin{aligned} &\mathbb{P}_n\{\xi_n \colon l_n \leq x | \mathcal{F}_{n-1}\} \\ &\quad = \frac{1}{\nu\{C_0\}}\, \nu\left\{\xi_n \colon f_1(\xi_n - \alpha\xi_{n-1}) \leq \frac{x}{\nu\{C_0\}}\right\} \\ &\quad = \frac{1}{\nu\{C_0\}}\, \nu\left\{w \colon f_1(w) \leq \frac{x}{\nu\{C_0\}}\right\} \end{aligned} \tag{12}$$

with the last quantity being a function only of x.

This result can be easily extended to more complicated dependency structures before and after the change. Specifically, consider processes of the form

$$\xi_n = g_n^i(\xi_{n-1}, \cdots, \xi_1) + w_n, \qquad i = 0, 1 \tag{13}$$

where $g_n^i(\xi_{n-1}, \cdots, \xi_1),\ i = 0, 1$ are nonlinear functions with the superscript "0" referring to the data model before the change and the superscript "1" after the change. The process w_n, as before, is i.i.d. and uniformly distributed before the change and has a density $f_1(w)$ after the change. Care must be taken to ensure that, for given $\xi_{n-1}, \cdots, \xi_1$, the points ξ_n that are accessible through (13) using the model before the change constitute a superset of the corresponding points that are accessible through the alternative model. This is necessary to guarantee the absolute continuity of the conditional measures described in the Introduction.

A final example in this class is when relation (13) applies to $\{\xi_n\}$ with the same function $g_n(\xi_{n-1}, \cdots, \xi_1)$ before and after the change and $\{w_n\}$ is i.i.d. with a different distribution (not necessarily one of the distributions being uniform). However, this is a case where the problem can be easily seen to reduce to an equivalent problem of testing i.i.d. processes since $\{w_n\}$ with $w_n = \xi_n - g_n(\xi_{n-1}, \cdots, \xi_1)$ has exactly the required i.i.d. statistics.

C. Processes Evolving on a Circle

The third and final example where key condition (2) can be valid is when we have processes evolving on a circle. Specifically, consider a circle of unit radius and let $\xi_n \in [0, 2\pi)$ denote the position of a point on the circle. Let us assume that, before the change, process $\{\xi_n\}$ is i.i.d. and uniformly distributed on the circle whereas after the change it satisfies the following random (on the circle) walk:

$$\xi_n = g(\xi_{n-1} + w_n) \tag{14}$$

with $\{w_n\}$ i.i.d. and with density $f_1(w)$. The function $g(\xi)$ is periodic with period 2π and defined as

$$g(\xi) = \xi - 2k\pi, \qquad \text{for } 2k\pi \leq \xi < 2(k+1)\pi \quad k = 0, \pm 1, \pm 2, \cdots. \tag{15}$$

If $h_1(\xi_n | \xi_{n-1})$ denotes the transition density after the change then for $\xi_n, \xi_{n-1} \in [0, 2\pi)$ we have

$$h_1(\xi_n | \xi_{n-1}) = \sum_{k=-\infty}^{\infty} f_1(\xi_n - \xi_{n-1} + 2k\pi) \tag{16}$$

where $\xi_n - \xi_{n-1} + 2k\pi$ are the possible jumps that can lead state ξ_{n-1} to ξ_n, consequently,

$$l_n = 2\pi \sum_{k=-\infty}^{\infty} f_1(\xi_n - \xi_{n-1} + 2k\pi). \tag{17}$$

Since l_n is a function of the difference $\xi_n - \xi_{n-1}$ it is easy to see that (2) is satisfied.

As pointed out by Prof. Ritov (Hebrew University of Jerusalem), for this example, there exists an alternative way to generate i.i.d. random variables. If we define $v_n = g(\xi_n - \xi_{n-1})$, with $g(\xi)$ introduced in (15), then one can show that $\{v_n\}$ is i.i.d. having a uniform density before the change and a density equal to $\Sigma_{k=-\infty}^{\infty} f_1(v + 2k\pi)$ after the change. The resulting l_n can be easily seen to be identical to the one defined in (17).

The above result can be extended to the case where, before the change, $\{\xi_n\}$ is uniformly distributed on the circle and after the change it satisfies

$$\xi_n = g(u_n(\xi_{n-1}, \cdots, \xi_1) + w_n) \tag{18}$$

with $g(\xi)$ as in (15) and $\{u_n(\xi_{n-1}, \cdots, \xi_1)\}$ any sequence of nonlinear functions.

IV. Conclusion

A simple observation that applies to all known stopping times that optimally solve the change detection and the hypotheses testing problem for the i.i.d. case, extends their optimality property to a class of processes that can have strong dependency structures. Several examples were shown to achieve optimal solutions through the use of these popular stopping times.

References

[1] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes, Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] E. Carlstein, H. Muller, and D. Seigmund, Eds., *Change-Point Problems*. Hayward, CA: Inst. Math. Stat., 1994.

[3] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Stat.*, vol. 42, pp. 1897–1908, 1976.

[4] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, pp. 1379–1387, 1986.

[5] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[6] M. Pollak, "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 13, pp. 206–227, 1985.

[7] M. Pollak and D. Siegmund, "Approximations to the expected sample size of certain sequential tests," *Ann. Statist.*, vol 6, pp. 1267–1282, 1975.

[8] Y. Ritov, "Decision theoretic optimality of the CUSUM procedure," *Ann. Satist.*, vol. 18, pp. 1464–1469, 1990.

[9] A. N. Shiryayev, "On optimal methods in earliest detection problems," *Theor. Probl. Appl.*, vol. 8, pp. 26–51, 1963.

[10] ——, *Optimal Stopping Rules*. New York: Springer-Verlag, 1978.

[11] A. Wald, *Sequential Analysis*. New York: Wiley, 1947.

[12] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Ann. Math. Statist*., vol. 19, pp. 326–339, 1948.

[13] B. Yakir, "Optimal detection of a change in distribution when the observations form a Markov chain with a finite state space," in *Change-Point Problems*, E. Carlstein, H. Muller, and D. Seigmund, Eds. Hayward, CA: Inst. Math. Stat., 1994.

[14] ——, "A note on optimal detection of a change in distribution," *Ann. Statist.*, vol. 25, no. 5, pp. 2117–2126, Oct. 1997.

On the Consistency of Minimum Complexity Nonparametric Estimation

Zhiyi Chi and Stuart Geman

***Abstract*— Nonparametric estimation is usually inconsistent without some form of regularization. One way to impose regularity is through a prior measure. Barron and Cover [1], [2] have shown that complexity-based prior measures can insure consistency, at least when restricted to countable dense subsets of the infinite-dimensional parameter (i.e., function) space. Strangely, however, these results are independent of the actual complexity assignment: the same results hold under an arbitrary permutation of the match-up of complexities to functions. We will show that this phenomenon is related to the weakness of the convergence measures used. Stronger convergence can only be achieved through complexity measures that relate to the actual behavior of the functions.**

***Index Terms*—Consistency, minimum complexity estimation, minimum description length, nonparametric estimation.**

I. Introduction

Maximum-likelihood, least squares, and other estimation techniques are generally inconsistent for nonparametric (infinite-dimensional) problems. Some variety of regularization is needed. An appealing and principled approach is to base regularization on complexity: Define an encoding of the (infinite-dimensional) parameter, and adopt codelength as a penalty. Barron and Cover [1], [2] have shown how to make this work. They get consistent estimation for densities and regressions, as well as some convergence-rate bounds, by constructing complexity-based penalty terms for maximum-likelihood and least squares estimators.

Can we cite the results of Barron and Cover as an argument for complexity-based regularization (or, equivalently, for complexity-based priors)? Apparently not: The results are independent of the particular assignment of complexities. Specifically, the results are unchanged by an arbitrary permutation of the matching of complexities to parameters.

Of course there are many ways to define convergence of functions. We will show here that the surprising indifference of convergence results to complexity assignments is in fact related to the convergence measures used. Stronger convergence requires a stronger tie between the parameters (functions) and their complexity measures.

Section II is a review of some Barron and Cover results. Then some new results about consistency for nonparametric regression are presented in Section III. (Proofs are in the Appendix.) Taken together, the results of Section III establish the principle that stronger types of convergence are sensitive to the particulars of the complexity assignment. We work here with regression, but the situation is analogous in density estimation.

Our results are about consistency only. The important practical issue of relating complexity measures to *rates* of convergence remains open.

II. Complexity-Based Priors

Barron and Cover [1] have shown that the problem of estimating a density nonparametrically can be solved using a complexity-based prior by limiting the prior to a countably-dense subset of the space of densities. More specifically, given a sequence of countable sets of densities Γ_n, and numbers $L_n(q)$ for densities q in Γ_n, let $\Gamma = \cup_n \Gamma_n$. Set $L_n(q) = \infty$ for q not in Γ_n. For independent random variables $X_1, X_2, \cdots, X_n$ drawn from an unknown probability density function p, a minimum complexity density estimator $\hat{p}_n$ is defined as a density achieving the following minimization:

$$\min_{q\in\Gamma_n}\left(L_n(q) - \sum_{i=1}^{n} \log\, q(X_i)\right).$$

If we think of $L_n(q)$ as the description length of the density q, then the minimization is over total description length—accounting for both the density and the data. Barron and Cover showed that if L_n satisfies the summability condition

$$\sup_n \sum_{q\in\Gamma_n} 2^{-L_n(q)} < +\infty$$

and the growth restriction

$$\limsup_n \frac{L_n(q)}{n} = 0, \qquad \text{for every}\, q \in \Gamma \tag{1}$$

then for each measurable set S

$$\lim_{n\to\infty} \hat{P}_n(S) = P(S) \quad \text{with probability one}$$

Manuscript received December 9, 1996; revised October 20, 1997. This work was supported by the Army Research Office under Contract DAAL03-92-G-0115, the National Science Foundation, under Grant DMS-9217655, and the Office of Naval Research, under Grant N00014-96-1-0647.

The authors are with the Division of Applied Mathematics, Brown University, Providence, RI 02912 USA.

Publisher Item Identifier S 0018-9448(98)04795-6.