on the design, analysis, and usage of vector quantization techniques in several application domains.

Another applicable area for BLFs is image processing: the paper by Li *et al.* [19] studies bounds on asymptotic performance of vector quantizers with perceptual distortion measure for which BLFs are natural candidates.

VI. Conclusion

This correspondence provides necessary and sufficient conditions for loss functions under which the conditional expectation is the unique optimal predictor. Beyond its mathematical interest, the expansion from the $\mathbb{L}^2$-loss function to the general class of BLFs has its own distinctive value. In areas such as image and speech codings where the $\mathbb{L}^2$-loss function is no longer an appropriate or even meaningful measure of error (as was pointed out in [20]), other functions such as the Kullback–Liebler (KL) divergence or the Itakura–Saito distance (see Table I) play a dominant role. Our findings may serve as a mathematical justification for the adoption of these loss functions.

Finally, as was alluded earlier, the stronger regularity condition for the high-dimensional case (Theorem 4) is used in a crucial way to verify the compatibility condition (11), which seems almost necessary for solving the system of (10). It will be interesting to see if the regularity condition can be relaxed.

References

[1] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed. San Diego, CA: Academic, 1974.

[2] O. Knill, "Probability," Course notes from Caifornia Institute of Technology, Pasadena, CA, 1994.

[3] D. Williams, *Probability With Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[4] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer-Verlag, 1998.

[5] K. B. Athreya, "Prediction Under Convex Loss," Dept. Mathematics and Statistics, Iowa State Univ., Ames, IA, Tech. Rep. 99-2, 1999.

[6] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. Phys.*, vol. 7, pp. 200–217, 1967.

[7] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, ser. SIAM Classics in Applied Mathematics. Philadelphia, PA: SIAM, 1999.

[8] R. T. Rockafellar, *Convex Analysis*, ser. Princeton Landmarks in Mathematics. Princeton, NJ: Princeton Univ. Press, 1970.

[9] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," in *Proc. SIAM Int. Conf. Data Mining*, 2004, pp. 234–245.

[10] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 3rd ed. New York: Springer-Verlag, 2001.

[11] C. H. Edwards, *Advanced Calculus of Several Variables*. Mineola, NY: Dover, 1995.

[12] H. L. V. Trees, *Detection, Estimation and Modulation Theory (Part I)*. New York: Wiley, 1968.

[13] Y. Censor and S. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*. London, U.K.: Oxford Univ. Press, 1998.

[14] I. Csiszár, "Why least sqaures and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Annals Stat.*, vol. 19, no. 4, pp. 2032–2066, 1991.

[15] A. Ben-Tal, A. Charnes, and M. Teboulle, "Entropic means," *J. Math. Anal. Appl.*, vol. 139, pp. 537–551, 1989.

[16] I. Csisár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar*, vol. 2, pp. 299–318, 1967.

[17] I. Csiszár, "Generalized projections for nonnegative functions," *Acta Mathematica Hungarica*, vol. 68, no. 1–2, pp. 161–185, 1995.

[18] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1991.

[19] J. Li, N. Chaddha, and R. M. Gray, "Asymptotic performance of vector quantizers with a perceptual distortion measure," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1082–1091, 1999.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.

Information Bounds and Quickest Change Detection in Decentralized Decision Systems

Yajun Mei

***Abstract*—The quickest change detection problem is studied in decentralized decision systems, where a set of sensors receive independent observations and send summary messages to the fusion center, which makes a final decision. In the system where the sensors do not have access to their past observations, the previously conjectured asymptotic optimality of a procedure with a monotone likelihood ratio quantizer (MLRQ) is proved. In the case of additive Gaussian sensor noise, if the signal-to-noise ratios (SNR) at some sensors are sufficiently high, this procedure can perform as well as the optimal centralized procedure that has access to all the sensor observations. Even if all SNRs are low, its detection delay will be at most $\pi/2 - 1 \approx 57\%$ larger than that of the optimal centralized procedure. Next, in the system where the sensors have full access to their past observations, the first asymptotically optimal procedure in the literature is developed. Surprisingly, the procedure has the same asymptotic performance as the optimal centralized procedure, although it may perform poorly in some practical situations because of slow asymptotic convergence. Finally, it is shown that neither past message information nor the feedback from the fusion center improves the asymptotic performance in the simplest model.**

***Index Terms*—Asymptotic optimality, CUSUM, multisensor, quantization, sensor networks, sequential detection.**

I. Introduction

The problem of quickest change detection has a variety of applications, including industrial quality control, reliability, fault detection, and signal detection. The classical or centralized version of this problem, where all observations are available at a single central location, is a well-developed area (see, e.g., [1], [7], and [17]). Recently, this problem has been applied in decentralized or distributed decision systems, which have many important applications, including multisensor data fusion, mobile and wireless communication, surveillance systems, and distributed detection.

Fig. 1 illustrates the general setting of decentralized decision systems. In such a system, at time n, each of a set of L sensors S_l receives an observation $X_{l,n}$ and then sends a sensor message $U_{l,n}$ to a central processor, called the *fusion center*, which makes a final decision when observations are stopped. In order to reduce cost and increase reliability, it is required that the sensor messages belong to a finite alphabet

Manuscript received November 21, 2002; revised November 10, 2004. This work was supported in part by the National Institutes of Health under Grant R01 AI055343. The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004.

The author was with the Department of Mathematics, California Institute of Technology, Pasadena, CA USA. He is now with the Department of Biostatistics, Fred Hutchinson Cancer Research Center, Seattle, WA 98109 USA (e-mail: ymei@fhcrc.org).

Communicated by A. Kavčić, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2005.850159

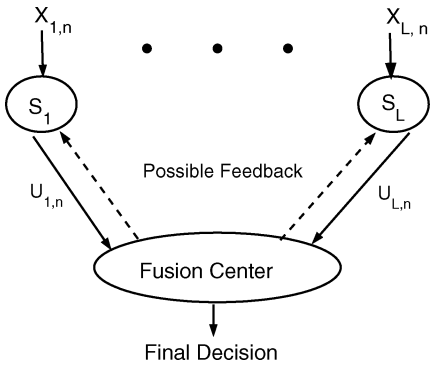

Fig. 1. General setting for decentralized decision systems.

(perhaps binary). This limitation is dictated in practice by the need for data compression and limitations of communication bandwidth.

In [23] and [25], the authors considered two different scenarios of decentralized decision systems, depending on how local information is used at the sensors. One scenario is the system with limited local memory, where the sensors do not have access to their past observations. This scenario has the following three possible cases, which correspond to Cases A, C, and E in [23] and [25].

Case 1 A system with neither feedback from the fusion center nor local memory, as follows:

$$U_{l,n} = \phi_{l,n}(X_{l,n}). \tag{1}$$

Case 2 A system with no feedback and local memory restricted to past sensor messages, as follows:

$$U_{l,n} = \phi_{l,n}\Big(X_{l,n}; U_{l,[1,n-1]}\Big) \tag{2}$$

where $U_{l,[1,n-1]} = (U_{l,1}, U_{l,2}, \ldots, U_{l,n-1})$.

Case 3 A system with full feedback and local memory restricted to past sensor messages, as follows:

$$U_{l,n} = \phi_{l,n}\Big(X_{l,n}; U_{1,[1,n-1]}, U_{2,[1,n-1]}, \ldots, U_{L,[1,n-1]}\Big). \tag{3}$$

The other scenario is the system with full local memory, where the sensors have full access to their past observations. There are two possible cases, which correspond to Cases B and D in [23] and [25].

Case 4 A system with no feedback and full local memory, as follows:

$$U_{l,n} = \phi_{l,n}(X_{l,[1,n]}), \tag{4}$$

where $X_{l,[1,n]} = (X_{l,1}, X_{l,2}, \ldots, X_{l,n})$.

Case 5 A system with full feedback and full local memory, as follows:

$$U_{l,n} = \phi_{l,n}\Big(X_{l,[1,n]}; U_{1,[1,n-1]}, U_{2,[1,n-1]}, \ldots, U_{L,[1,n-1]}\Big). \tag{5}$$

In decentralized quickest change detection problems, it is assumed that at some unknown time ν, the distributions of the sensor observations $X_{l,n}$ change abruptly and simultaneously at all sensors. The goal is to detect the change as soon as possible over all possible protocols for generating sensor messages and over all possible decision rules at the fusion center, under a restriction on the frequency of false alarms.

As in the classical or centralized quickest change detection problem, there are two standard mathematical formulations. The first one is a Bayesian formulation, due to Shiryayev [19], in which the change-point ν is assumed to have a known prior distribution. It is well known) [24], [25]) that Bayesian formulations prove to be intractable and the dynamic programming arguments cannot be used except in the special case specified in (5), where the Bayesian solution [24] is too complex to implement.

The second is a minimax formulation, proposed by Lorden [11], in which the change-point ν is assumed to be unknown (possibly ∞) but nonrandom. References [2] and [21] used this approach to study the simplest case specified in (1), but both have restrictions on the class of sensor message protocols.

In this correspondence, we use the second of these formulations to develop an asymptotic theory of decentralized quickest change detection problems, giving in both scenarios procedures that are asymptotically optimal and easy to implement. It is worthwhile highlighting that our asymptotically optimal procedures do not use past message information, and hence past message information (or the feedback from the fusion center) does not improve asymptotic performance.

Throughout this correspondence, we make two assumptions, which are standard.

(A1) The sensor observations are independent over time as well as from sensor to sensor.

(A2) The densities of the sensor observations are either $f_1, \ldots, f_L$ or $g_1, \ldots, g_L$, where the f's and g's are given. For each $1 \le l \le L$, the Kullback–Leibler information number (or relative entropy)

$$I(g_l, f_l) = \int \log\left(\frac{g_l(x)}{f_l(x)}\right) g_l(x)\,dx \tag{6}$$

is finite and positive, and

$$\int \left(\log \frac{g_l(x)}{f_l(x)}\right)^2 g_l(x)\,dx < \infty. \tag{7}$$

In Section II, we provide a formal mathematical formulation of the problem and introduce some notations. In Section III, under a condition on second moments, we prove that a procedure with a monotone likelihood ratio quantizer (MLRQ) is asymptotically optimal in the system with limited local memory. We also establish sufficient conditions for our theorems to be applied. Section IV develops asymptotic theory in the system with full local memory and offers asymptotic optimal procedures that are easy to implement. In Section V, we compare these asymptotically optimal decentralized procedures with the optimal centralized procedure that has access to all the sensor observations. Section VI gives simulation results for several illustrative examples. The proofs of all theorems are given in the Appendix .

II. Problem Formulation and Notation

Suppose there are L sensors in a system. At time n, an observation $X_{l,n}$ is made at each sensor S_l. Assume that at some unknown (possibly ∞) time ν, the density function of the sensor observations $\{X_{l,n}\}$ changes simultaneously for all $1 \le l \le L$ from f_l to g_l. That is, for each $1 \le l \le L$, the observations at sensor S_l, $X_{l,1}, X_{l,2}, \ldots$ are independent random variables such that $X_{l,1}, X_{l,2}, \ldots, X_{l,\nu-1}$ are independent identically distributed (i.i.d.) with density f_l and $X_{l,\nu}, X_{l,\nu+1}, \ldots$ are i.i.d. with density g_l. Furthermore, it is assumed

that the observations are independent from sensor to sensor. Denote by P_ν and E_ν the probability measure and expectation when the change occurs at time ν, and denote the same by P_∞ and E_∞ when there is no change.

Based on the information available at S_l at time n, a message $U_{l,n}$, specified in (1)–(5), is chosen from a finite alphabet and is sent to a fusion center. Without loss of generality, we assume that $U_{l,n}$ takes a value in $\{0, 1, \ldots, D_l - 1\}$. The fusion center uses the stream of messages from the sensors as inputs to make a decision whether or not a change has occurred.

Mathematically, the fusion center decision rule is defined as a stopping time τ with respect to $\{(U_{1,n}, U_{2,n}, \ldots, U_{L,n})\}_{n\geq 1}$. The interpretation of τ is that, when $\tau = n$, we stop taking observations at time n and declare that a change has occurred somewhere in the first n observations.

For each choice of sensor message functions and fusion center decision rule, a reasonable measure of quickness of detection is the following "worst case" detection delay defined in Lorden [11]:

$$\overline{E}_1(\tau) = \sup_{\nu\geq 1}\left(\operatorname{ess\,sup} E_\nu\left[(\tau-\nu+1)^+ \middle| X_{1,[1,\nu-1]}, \ldots, X_{L,[1,\nu-1]}\right]\right).$$

The desire to have small $\overline{E}_1(\tau)$ must, of course, be balanced against the need to have a controlled frequency of false alarms. In other words, when no change occurs, τ should be large, hopefully infinite. However, Lorden [11] showed that if $\overline{E}_1(\tau)$ is finite, then $E_\infty \tau$ is finite, which implies $P_\infty(\tau < \infty) = 1$. Thus, we will have a false alarm with probability 1 when there is no change. An appropriate measurement of false alarms, therefore, is $E_\infty \tau$, the mean time until a false alarm. Imagining repeated application of such procedures, practitioners refer to the frequency of false alarms as $1/E_\infty \tau$ and the mean time between false alarms as $E_\infty \tau$.

Our problem can then be stated as follows: Design the sensors' message function $\phi_{l,n}$ and seek a stopping time τ at the fusion center that minimizes $\overline{E}_1(\tau)$ subject to

$$E_\infty \tau \geq \gamma \tag{8}$$

where γ is a given, fixed lower bound.

The worst case detection delay $\overline{E}_1(\tau)$ can be replaced by the "average" detection delay, proposed by Shiryayev [20] and Pollak [15]

$$\sup_{\nu\geq 1} E_\nu(\tau - \nu | \tau \geq \nu).$$

Although the worst-case detection delay is always greater than the average detection delay, they are asymptotically equivalent. Either one can be used in our theorems.

It is well known ([13]) that the (exactly) optimal solutions for this problem in the centralized version are Page's cumulative sum (CUSUM) procedures, defined by the stopping times

$$T(a) = \inf\left\{n : \boldsymbol{W}_n \geq a\right\} \tag{9}$$

where the CUSUM statistic

$$\boldsymbol{W}_n = \max_{1\leq k\leq n} \sum_{i=k}^{n}\left(\sum_{l=1}^{L} \log \frac{g_l(X_{l,i})}{f_l(X_{l,i})}\right)$$

which can be calculated recursively as

$$\boldsymbol{W}_n = \max\left(\boldsymbol{W}_{n-1},\ 0\right) + \sum_{l=1}^{L} \log \frac{g_l(X_{l,n})}{f_l(X_{l,n})} \tag{10}$$

for $n \geq 1$ and $\boldsymbol{W}_0 = 0$. In the literature, $T(a)$ is also usually defined as the first n for which $\max(\boldsymbol{W}_n, 0) \geq a$. These two definitions are equivalent if the threshold $a > 0$, but there is a difference if $a \leq 0$ (see also [13]).

Unfortunately, in decentralized decision systems, it is nearly impossible to find exactly optimal solutions (for some special cases, see [24]), and only "asymptotic optimality" results seem to be working. In the asymptotic optimality approach, we typically first construct an asymptotic lower bound of $\overline{E}_1\tau$ as γ goes to ∞. Then, we show that a given class of procedures attains the lower bound asymptotically. We will establish asymptotic optimality theorems for both scenarios of decentralized decision systems: limited local memory [specified in (1)–(3)] and full local memory [specified in (4) and (5)].

We now introduce some notations. Let D be a positive integer. Consider a random variable Y whose density function is either f or g with respect to some σ-finite measure, and assume that the Kullback–Leibler information number $I(g, f)$ is finite. For a (*deterministic* or *random*) measurable function ϕ from the range of Y to a finite alphabet of size D, say $\{0, 1, \ldots, D-1\}$, denote by f_ϕ and g_ϕ, respectively, the probability mass function of $\phi(Y)$ when the density of Y is f and g. Let

$$Z_\phi = \log \frac{g_\phi(\phi(Y))}{f_\phi(\phi(Y))}$$

and define

$$I_D(g, f) = \sup_\phi E_g Z_\phi \tag{11}$$

and

$$V_D(g, f) = \sup_\phi E_g(Z_\phi)^2. \tag{12}$$

It is well known [22] that $I_D(g, f) \leq I(g, f)$, i.e., that reduction of the data from Y to $\phi(Y)$ cannot increase the information. A more detailed analysis between $I_D(g, f)$ and $I(g, f)$ is provided in Section V. Tsitsiklis [22] showed that the supremum $I_D(g, f)$ is achieved by an MLRQ φ of the form

$$\varphi(Y) = d \quad \text{if and only if} \quad \lambda_d \leq \frac{g(Y)}{f(Y)} < \lambda_{d+1}$$

where $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{D-1} \leq \lambda_D = \infty$ are constants. These optimal MLRQs are not easily calculated, but we follow the standard practice in the literature of developing procedures that assume sensor messages are constructed optimally in the sensor. Some of our theorems assume that $V_D(g, f) < \infty$. A sufficient condition for finiteness of $V_2(g, f)$ is given in Section III.

Using these notations, define the information numbers

$$I_{\boldsymbol{D}} = \sum_{l=1}^{L} I_{D_l}(g_l, f_l) \tag{13}$$

where $\boldsymbol{D} = (D_1, D_2, \ldots, D_L)$, and

$$I_{\text{tot}} = \sum_{l=1}^{L} I(g_l, f_l). \tag{14}$$

These two information numbers are key to our theorems.

III. Limited Local Memory

A. Page's CUSUM Procedure With the MLRQ

For the decentralized decision system with limited local memory, specified in (1)–(3), the following procedure $N(a)$ has been studied in the literature:

Each sensor S_l uses the optimal MLRQ φ_l. Namely

$$U_{l,n} = \varphi_l(X_{l,n}) = d \text{ if and only if } \lambda_{l,d} \leq \frac{g_l(X_{l,n})}{f_l(X_{l,n})} < \lambda_{l,d+1}$$

where $0 = \lambda_{l,0} \le \lambda_{l,1} \le \cdots \le \lambda_{l,D_l-1} \le \lambda_{l,D_l} = \infty$ are optimally chosen in the sense that the Kullback–Leibler information number $I(g_{\varphi,l}, f_{\varphi,l})$ achieves the supremum $I_{D_l}(g_l, f_l)$. Here, $f_{\varphi,l}$ and $g_{\varphi,l}$ are the probability mass functions induced on $U_{l,n}$ when the observations $X_{l,n}$ are distributed as f_l and g_l, respectively.

Based on the i.i.d. vector observations $\boldsymbol{U}_n = (U_{1,n}, \ldots, U_{L,n})$, the fusion center then uses Page's CUSUM procedure with log-likelihood ratio boundary a to detect whether or not a change has occurred, i.e., the stopping time $N(a)$ is given by

$$N(a) = \inf\{n : \hat{\boldsymbol{W}}_n \ge a\} \tag{15}$$

where $\hat{\boldsymbol{W}}_0 = 0$ and for $n = 1, 2, \ldots$,

$$\hat{\boldsymbol{W}}_n = \max\left(\hat{\boldsymbol{W}}_{n-1},\ 0\right) + \sum_{l=1}^{L} \log \frac{g_{\varphi,l}(U_{l,n})}{f_{\varphi,l}(U_{l,n})}.$$

It was shown in [2] that $N(a)$ is optimal in the sense that at each sensor, the MLRQ φ is optimized, i.e., maximizes the Kullback–Leibler information number $I(g_\varphi, f_\varphi)$. Later [21] proved the asymptotic optimality property of $N(a)$ in the simplest case specified in (1) under the restriction that the sensor message functions $\{\phi_1, \ldots, \phi_L\}$ satisfy the following "stationary" condition: For all $\nu = 1, 2, \ldots$, as n goes to ∞, $n^{-1}\sum_{i=\nu}^{\nu+n}\sum_{l=1}^{L} Z_{l,i}$ converges in probability under P_ν to some positive constant number, where

$$Z_{l,i} = \log\left(g_{\phi,l}(U_{l,i})/f_{\phi,l}(U_{l,i})\right).$$

Reference [24] conjectured that $N(a)$ is asymptotically optimal in the special case specified in (5), because numerical simulations illustrate that it has performance similar to the Bayesian solutions. In the next subsection, we will show that under a condition on second moments, $N(a)$ is asymptotically optimal without any restriction on the sensors' message functions or the fusion center decision rule in the system with limited local memory.

B. Asymptotic Optimality of $N(a)$

We begin our analysis by studying the performance of the procedure $N(a)$. Observe that $N(a)$, defined in (15), is Page's CUSUM procedure so that by applying the standard bounds [17], we get the following.

Lemma 1:

$$E_\infty N(a) \ge e^a$$

and as $a \to \infty$

$$\overline{E}_1 N(a) \le \frac{a}{I_{\boldsymbol{D}}} + O(1).$$

The following theorem is of fundamental importance for proving asymptotic optimality of $N(a)$. It establishes the asymptotic lower bounds for the detection delays of any procedures in the system with limited local memory.

Theorem 1: Assume $V_{D_l}(g_l, f_l)$, defined in (12), is finite for all $1 \le l \le L$. If $\{\tau(\gamma)\}$ is a family of procedures in the system with limited local memory satisfying (8), then

$$\overline{E}_1 \tau(\gamma) \ge (1 + o(1))\frac{\log \gamma}{I_{\boldsymbol{D}}} \tag{16}$$

as $\gamma \to \infty$, where $I_{\boldsymbol{D}}$ is defined in (13).

Now, we can summarize our results on the asymptotic optimality of the procedure $N(a)$ as follows.

Corollary 1: For $\gamma > 1$, let $a = \log \gamma$, then $N(a)$ satisfies (8) and

$$\overline{E}_1 N(a) \le \frac{\log \gamma}{I_{\boldsymbol{D}}} + O(1)$$

so that under the assumption of finiteness of $V_{D_l}(g_l, f_l)$ for all $1 \le l \le L$, the procedure $N(a)$ asymptotically minimizes the detection delay $\overline{E}_1 N(a)$ as $\gamma \to \infty$ in the system with limited local memory.

Note that reference [21] established a result similar to (16) in the simplest case specified in (1) under a restriction on the sensor message functions. Theorem 1 provides different sufficient conditions under which the asymptotic lower bounds (16) could be established. Our sufficient conditions are new and perhaps the most useful, since they do not impose any restrictions on the sensors' message functions or the fusion center decision rules. Moreover, they also allow us to obtain the asymptotic optimality property of $N(a)$ in all three cases of the system with limited local memory.

C. Sufficient Conditions

In Theorem 1, we assume $V_D(g, f) < \infty$, which is usually not easy to verify. The following theorem and its corollary give some sufficient conditions to verify it when $D = 2$.

Theorem 2: Suppose $f(y)$ and $g(y)$ are two densities such that

$$E_g\left(\log \frac{g(Y)}{f(Y)}\right)^2 = \int \left(\log \frac{g(y)}{f(y)}\right)^2 g(y)dy < \infty.$$

Define

$$A(t) = P_f\left(\frac{g(Y)}{f(Y)} > t\right), \quad B(t) = P_g\left(\frac{g(Y)}{f(Y)} > t\right).$$

Assume $A(t)$ and $B(t)$ are continuous functions of t on $(0, \infty)$, and take values 0 and 1 for the same t. Moreover, assume that

$$\limsup_{t\to\infty} \sqrt{B(t)}\left|\log A(t)\right| < \infty \tag{17}$$

and

$$\limsup_{t\to 0} \sqrt{1 - A(t)}\left|\log(1 - B(t))\right| < \infty \tag{18}$$

where $\sqrt{0}|\log 0|$ is interpreted as 0. Then, $V_2(g, f) < \infty$.

Corollary 2: Suppose the distribution of the random variable Y belongs to a one-parameter exponential family having the continuous densities

$$f_\theta(y) = \exp\{\theta y - b(\theta)\}, \qquad -\infty < y < \infty,\ \theta \in \Omega$$

with respect to some σ-finite measure, where Ω is an open interval on the real line and $b(\theta)$ is twice differentiable with respect to θ. Let $F_\theta(y)$ denote the distribution function of Y. Consider $\theta_0 < \theta_1$ in Ω, and let $f_i = f_{\theta_i}$ and $F_i = F_{\theta_i}$ for $i = 0, 1$. Define $y_0 = \sup\{y : F_0(y) = 0\}$ and $y_1 = \inf\{y : F_1(y) = 1\}$. If

$$\lim_{y\to y_0} \frac{(F_0(y))^{3/2}}{f_0(y)} < \infty \quad \text{and} \quad \lim_{y\to y_1} \frac{(1 - F_1(y))^{3/2}}{f_1(y)} < \infty$$

then both $V_2(f_0, f_1)$ and $V_2(f_1, f_0)$ are finite.

Proof: Since $f_1(y)/f_0(y)$ is a monotonically increasing function of y, it suffices for $V_2(f_1, f_0) < \infty$ to show (17) and (18) hold for $A(t) = 1 - F_0(\log t)$ and $B(t) = 1 - F_1(\log t)$, which is straightforward using L'Hôpital's rule. The proof is identical for $V_2(f_0, f_1)$. □

It is easy to check that two Gaussian distributions with same variance satisfy the conditions in Corollary 2 and so do two exponential distributions. Therefore, if the sensors are restricted to send binary messages to the fusion center, and the prechange and postchange distributions at each sensor are two Gaussian distributions with same variance or two exponential distributions, then the procedure $N(a)$ is asymptotically optimal (over all possible sensor messages and all possible fusion center decision rules) in the system with limited local memory.

IV. Full Local Memory

It has been an open problem to find asymptotically optimal procedures (including both the sensor and fusion center decision rules) in the decentralized decision system with full local memory, specified in (4) and (5), although it is well-known [25] that Bayesian formulations become intractable. We will address this problem in this section.

To establish lower bounds for the detection delay in the system with full local memory is not difficult. By the optimality of Page's CUSUM procedures in the centralized version ([11], [13], [17]), we have the following.

Lemma 2: If $\{\tau(\gamma)\}$ is a family of procedures in the system with full local memory such that (8) holds, then as $\gamma \to \infty$

$$\overline{E}_1\tau(\gamma) \geq \frac{\log\gamma}{I_{\rm tot}} + O(1). \tag{19}$$

In the centralized version, the lower bounds (19) are sharp and can be achieved by Page's CUSUM procedure $T(a)$ defined in (9). Theorem 1 shows that these lower bounds are too crude in the system with limited local memory. However, it is not clear whether they are sharp in the system with full local memory. In other words, can we find procedures in the system with full local memory for which these bounds are achieved asymptotically? Since we expect to sacrifice some performance by quantizing the data locally instead of utilizing the complete data set at the fusion center, it is perhaps surprising that we give an affirmative answer by constructing such procedures.

A. The Structure of Procedures

For the system with full local memory, our proposed procedure $M(a)$ is as follows.

For each sensor S_l, one considers whether or not the CUSUM statistic

$$W_{l,n} = \max_{1\leq k\leq n}\sum_{i=k}^{n}\log\frac{g_l(X_{l,i})}{f_l(X_{l,i})} \tag{20}$$

exceeds the constant boundary $\pi_l a$, where

$$\pi_l = \frac{I(g_l,f_l)}{\sum_{l=1}^{L} I(g_l,f_l)} = \frac{I(g_l,f_l)}{I_{\rm tot}}. \tag{21}$$

That is, for each $l = 1,\ldots,L$, and $n = 1,2,\ldots$, define the sensor messages

$$U_{l,n} = \begin{cases} 1, & \text{if } W_{l,\,n} \geq \pi_l a \\ 0, & \text{otherwise.}\end{cases}$$

The fusion center then combines all these "sensor decisions" $U_{l,n}$ by using an AND rule, i.e., it stops and decides a change has occurred as soon as $U_{l,n} = 1$ for all $l = 1,2,\ldots,L$.

This stopping time $M(a)$ can be written as

$$M(a) = \inf\Big\{n \geq 1 : W_{l,n} \geq \pi_l a \ \text{ for all } \ l = 1,2,\ldots,L\Big\}. \tag{22}$$

It is easy to see that in single-sensor systems, our procedure $M(a)$ coincides with the optimal centralized procedure $T(a)$, defined in (9). Similar to $T(a)$, it is very convenient to implement $M(a)$ because the CUSUM statistic $W_{l,n}$ obeys the recursive relation

$$W_{l,n} = \max\Big\{W_{l,n-1},\,0\Big\} + \log\frac{g_l(X_{l,n})}{f_l(X_{l,n})}$$

where $W_{l,0} = 0$. However, unlike $T(a)$, our procedure $M(a)$ requires that each sensor shall continue sending the local messages to the fusion center even after the CUSUM statistic exceeds the local threshold. This essential feature can be seen from the following heuristic argument, which provides the motivation of $M(a)$.

Consider the optimal centralized procedure $T(a)$, defined in (9). If ν is the true change-point and $n-\nu$ is sufficiently large, then

$$\begin{aligned}\boldsymbol{W}_n &= \max_{1\leq k\leq n}\sum_{i=k}^{n}\left(\sum_{l=1}^{L}\log\frac{g_l(X_{l,i})}{f_l(X_{l,i})}\right)\\ &\approx \sum_{i=\nu}^{n}\left(\sum_{l=1}^{L}\log\frac{g_l(X_{l,i})}{f_l(X_{l,i})}\right)\end{aligned}$$

and

$$W_{l,n} = \max_{1\leq k\leq n}\sum_{i=k}^{n}\log\frac{g_l(X_{l,i})}{f_l(X_{l,i})} \approx \sum_{i=\nu}^{n}\log\frac{g_l(X_{l,i})}{f_l(X_{l,i})}.$$

Thus, $\boldsymbol{W}_n \approx \sum_{l=1}^{L} W_{l,n}$, and so under P_ν, the stopping rule of the optimal centralized procedure $T(a)$ is roughly equivalent to

$$\left\{\sum_{l=1}^{L} W_{l,n} \geq a\right\} \tag{23}$$

for sufficiently large a. Now, the strong law of large numbers implies that $(n-\nu)^{-1}W_{l,n} \to I(g_l,f_l)$ with probability 1, so the weight of $W_{l,n}$ in the sum is roughly $I(g_l,f_l)/(\sum_{l=1}^{L} I(g_l,f_l)) = \pi_l$. Thus, (23) can be approximated by $\{W_{l,n} \geq \pi_l a \text{ for all } 1 \leq l \leq L\}$, which is exactly the stopping rule of our procedure $M(a)$.

B. Asymptotic Optimality

The following theorem, whose proof is substantially complicated, establishes the asymptotic properties of our procedure $M(a)$ for large values of a.

Theorem 3: As $a \to \infty$

$$\overline{E}_1 M(a) \leq \frac{a}{I_{\rm tot}} + (C + o(1))\sqrt{\frac{a}{I_{\rm tot}}} \tag{24}$$

where $I_{\rm tot}$ is defined in (14), and $C > 0$ is a constant depending on L and the densities f_l and g_l. Furthermore, if we assume

$$\int g_l(x)\left|\log\frac{g_l(x)}{f_l(x)}\right|^3 dx < \infty \tag{25}$$

for each $1 \leq l \leq L$, then as $a \to \infty$

$$E_\infty M(a) \geq (1 + o(1))e^a. \tag{26}$$

Remark 1: Under additional reasonable conditions, it follows from nonlinear renewal theory that the smallest constant C in (24) is given by

$$C = E\max_{1\leq l\leq L}\left(\frac{\sigma_l}{I(g_l,f_l)}Z_l\right) \tag{27}$$

where $\sigma_l^2 = \mathrm{Var}_{g_l}\Big(\log(g_l(X)/f_l(X))\Big)$ and $Z_1, \ldots, Z_L$ are independent standard Gaussian variables. The proof is same as that of [4, Theorem 3.3] (Also see [3, Lemma 1].)

Remark 2: For each sensor, the mean time between false alarms is $\exp(\pi_l a)$. By the renewal property of the CUSUM statistics, the mean time between false alarms for the fusion center is of order $\prod_{l=1}^{L} \exp(\pi_l a) = \exp(a)$ since we continue sending local messages. (See the Appendix for the rigorous proof. As in [18], the key idea is Lemma 6 in the Appendix.)

Remark 3: Lemma 6 in the Appendix indicates that our procedure $M(a)$ has the same pleasant property as the procedure $N(a)$ in (15) and Page's CUSUM procedure $T(a)$ in (9): the mean time between false alarms is approximately exponentially distributed.

Remark 4: It is important to emphasize that in the definition of our procedure $M(a)$ in (22), we cannot replace the CUSUM statistics $W_{l,n}$ by the log-transformed Shiryayev–Roberts statistics

$$\log\Big(\sum_{k=1}^{n}\prod_{i=k}^{n}\Big(g_l(X_{l,i})/f_l(X_{l,i})\Big)\Big);$$

in that case, the mean time between false alarms is roughly $\exp((\max_{l=1}^{L}\pi_l)a)$, which is much smaller than $\exp(a)$ as $a \to \infty$.

Now, the asymptotic optimality of our procedure $M(a)$ follows at once from Theorem 3 and Lemma 2.

Corollary 3: There exists $a = \log\gamma + o(1)$ so that $M(a)$ satisfies (8) and

$$\overline{E}_1 M(a) \le \frac{\log\gamma}{I_{\mathrm{tot}}} + (C + o(1))\sqrt{\frac{\log\gamma}{I_{\mathrm{tot}}}}.$$

Thus, $M(a)$ minimizes the detection delay up to $O(\sqrt{\log\gamma})$ among all procedures in the system with full local memory satisfying (8).

V. Comparison of Three Procedures

In this section, we compare our asymptotically optimal decentralized procedures with the optimal centralized procedure. As in [2], for a decentralized procedure $\tau(\gamma)$ satisfying (8), define the decentralized penalty function (DPF)

$$\mathrm{DPF}_\tau(\gamma) = \frac{\overline{E}_1\tau(\gamma)}{n(\gamma)} - 1 \tag{28}$$

where $n(\gamma)$ is the detection delay of the optimal centralized procedure satisfying (8). Intuitively, DPF_τ can be thought of as a measure that reflects the relative performance degradation for using decentralized procedure τ instead of the optimal centralized procedure.

By Corollary 1 and relation (19), we immediately have Proposition 1.

Proposition 1: The DPF function of the procedure $N(a)$, defined in (15), is given by

$$\mathrm{DPF}_N(\gamma) = \frac{I_{\mathrm{tot}}}{I_{\boldsymbol{D}}} - 1 + O(\frac{1}{\log\gamma}). \tag{29}$$

It is, therefore, natural to study the relation between $I_{\boldsymbol{D}}$ and I_{tot}. By definition, it suffices to study the relation between $I_D(g,f)$ and $I(g,f)$ for a pair of densities (f,g). However, little research has been done on finding good lower bounds for $I_D(g,f)/I(g,f)$, although it is well known that the upper bound is 1. In the following, we study the special case of Gaussian distributions when $D = 2$. The idea can be easily extended to non-Gaussian distributions.

Proposition 2: Suppose $f(y)$ and $g(y)$ are two Gaussian distributions with respective mean μ_0 and μ_1 and same variance σ^2. Let $\rho = (\mu_1 - \mu_0)^2/(2\sigma^2)$ denote the signal-to-noise ratio (SNR), then

$$\liminf_{\rho\to 0}\frac{I_2(g,f)}{I(g,f)} \ge \frac{2}{\pi} \quad \text{and} \quad \lim_{\rho\to\infty}\frac{I_2(g,f)}{I(g,f)} = 1.$$

Proof: Without loss of generality, we assume $\mu_0 = 0$ and $\sigma = 1$. First note that $I(g,f) = \rho$ in this case. Next, since the likelihood ratio $g(y)/f(y)$ is a monotonically increasing function, the MLRQ can be written as

$$U = \begin{cases} 1, & Y \ge \lambda \\ 0, & \text{otherwise}. \end{cases}$$

Thus, the Kullback–Leibler information number for U is

$$r(\lambda) = h\Big(\Phi(\lambda - \mu_1), \Phi(\lambda)\Big) \tag{30}$$

where $\Phi(\cdot)$ is the distribution function of a standard Gaussian random variable and $h(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$. Now, let $\lambda = k\mu_1$. Fix k, and it is straightforward to show that

$$\frac{r(k\mu_1)}{I(g,f)} = \begin{cases} 2/\pi, & \text{as } \rho \to 0 \\ k^2 1\{0 < k < 1\} & \text{as } \rho \to \infty \end{cases}$$

where $1\{A\}$ is the indicator of the event A. The proposition follows at once from the fact $r(k) \le I_2(g,f) \le I(g,f)$ for any k. □

In the decentralized decision system with Gaussian sensor observations where the SNRs at *some* sensors are sufficiently high, we have $I_{\boldsymbol{D}} \approx I_{\mathrm{tot}}$ because those sensors with high SNRs will contribute most to I_{tot} and $I_{\boldsymbol{D}}$. Hence, the procedure $N(a)$ will perform as well as the optimal centralized procedure. Even if all SNRs are very low, the DPF function of $N(a)$ will be at most $\pi/2 - 1 \approx 57\%$ for large values of the mean time between false alarms. That is, the detection delay of $N(a)$ will be at most 57% larger than that of the optimal centralized procedure. In other words, the procedure $N(a)$ will take at most 57% more observations from the postchange distributions than the optimal centralized procedure. Moreover, the number of sensors does not have much effect on the DPF function of $N(a)$. Furthermore, Proposition 2 motivates us to conjecture that for Gaussian distributions, $I_2(g,f)/I(g,f)$ is an increasing function of SNR ρ with the range $[2/\pi, 1]$. We do not have a rigorous proof; however, numerical results support our conjecture.

Now, for the procedure $M(a)$ in the system with full local memory, by Corollary 3.

Proposition 3: The DPF function of the procedure $M(a)$, defined in (22), satisfies

$$\mathrm{DPF}_M(\gamma) \le (C + o(1))\frac{\sqrt{I_{\mathrm{tot}}}}{\sqrt{\log\gamma}} \tag{31}$$

where the constant C depends on L and the densities f_l and g_l.

It is easy to see that the DPF function of $M(a)$ is 0 as γ goes to ∞. That is, $M(a)$ can perform as well as the optimal centralized procedure in any systems if γ is sufficiently large. Unfortunately, the asymptotic convergence of $M(a)$ is so slow that $M(a)$ may perform very far from the optimum for realistic values of the mean time between false alarms in some systems. As an illustration, let us consider the symmetric Gaussian system where for each l, f_l, and g_l are Gaussian distributions with respective mean μ_0 and μ_1 and same variance σ^2. In this case, by (27)

$$C\sqrt{I_{\mathrm{tot}}} = \sqrt{2L}\cdot E(\max_{1\le l\le L} Z_l)$$

where $Z_1, \ldots, Z_L$ are independent standard Gaussian variables. Thus, the DPF function of $M(a)$ depends heavily on the number of sensors in this case. Using [4, Table I], we have

$$C\sqrt{I_{\rm tot}} = \begin{cases} 1.1284, & \text{if } L = 2 \\ 2.0730, & \text{if } L = 3 \\ 2.9115, & \text{if } L = 4 \\ 6.8815, & \text{if } L = 10. \end{cases}$$

For moderate values of γ, say $\gamma = 10^4$, we have $\sqrt{\log \gamma} \approx 3.03$, and so the right-hand side of (31) is roughly 37%, 68%, 96%, and 227%, respectively, if $L = 2, 3, 4$, and 10. This indicates that $M(a)$ may perform poorly for moderate values of γ in symmetric systems with multiple sensors. For example, when $L = 4$, the detection delay of $M(a)$ may be 96% larger than that of the optimal centralized procedure if $\gamma \approx 10^4$.

Finally, let us compare $M(a)$ with $N(a)$. While $M(a)$ has better asymptotic performance than $N(a)$, it is possible that $M(a)$ has worse performance than $N(a)$ in practical applications, especially when L, the number of sensors, is large but γ is only moderately large. To indicate this, note that the right-hand side of (31) could be larger than that of (29) if

$$\sqrt{\log \gamma} \le \frac{C\sqrt{I_{\rm tot}}}{I_{\rm tot}/I_{\boldsymbol{D}} - 1}. \tag{32}$$

Thus, if $C\sqrt{I_{\rm tot}}$ is large or $I_{\rm tot}/I_{\boldsymbol{D}}$ is small, then it is likely that $M(a)$ will perform worse than $N(a)$ for moderate values of γ. By (27), it is easy to see that if there are large number of sensors, then the value C will be very large, and so $M(a)$ can perform worse than $N(a)$. For instance, in the above symmetric Gaussian system with small SNRs, (32) becomes

$$\gamma \le \begin{cases} 50, & \text{if } L = 2 \\ 5.3 \times 10^5, & \text{if } L = 3 \\ 2.0 \times 10^{11}, & \text{if } L = 4 \\ 1.3 \times 10^{63}, & \text{if } L = 10. \end{cases}$$

Therefore, for moderate values of γ, say 10^4, it is likely that $M(a)$ will perform worse than $N(a)$ in the system with large number of sensors.

Observe that both of $N(a)$ and $M(a)$ do not use past message information or the feedback from the fusion center, but they are asymptotically optimal in the corresponding decentralized decision systems. This fact proves the following interesting result, part of which was conjectured in [24]:

Theorem 4: If all prechange and postchange distributions are completely specified and satisfy the conditions of Theorems 1 and 3, then neither past message information nor the feedback from the fusion center improves asymptotic performance in the decentralized decision systems specified in (1)–(5).

It should be pointed out that one of the underlying assumptions of this theorem is that the observations are independent from sensor to sensor. It is likely that past message information or the feedback will be more useful in practical applications where the observations are dependent or observation distributions are only partially specified.

VI. Numerical Results

In this section, we present a numerical illustration of the asymptotic theory of previous sections. Suppose there are L sensors each sending binary message to the fusion center, i.e., $D_l = 2$. Assume that the observations at sensor S_l are independent and identically distributed random variables with mean 0 and variance 1 before the change and with mean μ_l and variance 1 after the change. An interesting application of this model can be found in [21], where L geographically separated sensors are used to detect the appearance of a deterministic signal (or target), which is contaminated by additive white Gaussian noise at each sensor.

If $\mu_l > 0$, then the likelihood ratio at sensor S_l is a monotonically increasing function of the observation, and hence the MLRQ at each sensor S_l can be written as

$$U_{l,n} = \begin{cases} 1, & X_{l,n} \ge \lambda_l \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the Kullback–Leibler information number for $U_{l,n}$ is

$$r(\lambda_l) = h\Big(\Phi(\lambda_l - \mu_l), \Phi(\lambda_l)\Big)$$

where $\Phi(\cdot)$ and $h(a, b)$ are defined as in (30). Since the function $r(\lambda_l)$ has a unique maximum value over $[0, \infty]$, it is easy to find the optimal λ_l numerically. For example, if $\mu_l = 0.2$ or 1, then the optimal thresholds λ_l are 0.1584 and 0.7941, respectively, and the corresponding optimal Kullback–Leibler information numbers $r(\lambda_l)$ are 0.01273 and 0.3186, respectively. Note that in these situations, $I_2(g_l, f_l)/I(g_l, f_l)$ is close to $2/\pi$ since the Kullback–Leibler information number $I(g_l, f_l) = \mu_l^2/2$.

As an illustration, six cases are considered.

- Case 1 Two nonidentical sensors: $L = 2$, $\mu_1 = 0.2$, and $\mu_2 = 1$.
- Case 2 Two identical sensors: $L = 2$ and $\mu_1 = \mu_2 = 1$.
- Case 3 Three nonidentical sensors: $L = 3$, $\mu_1 = \mu_2 = 0.2$, and $\mu_3 = 1$.
- Case 4 Three identical sensors: $L = 3$ and $\mu_l = 1$ for $l = 1, 2, 3$.
- Case 5 Ten nonidentical sensors: $L = 10$, $\mu_l = 1$ if $l = 1, 2, 3$, and $\mu_l = 0.2$ if $4 \le l \le 10$.
- Case 6 Ten identical sensors: $L = 10$ and $\mu_l = 0.2$ for all $1 \le l \le 10$.

In each case, we compare three asymptotically optimal procedures, as follows:

i) $N(a)$, defined by (15) in the system with limited local memory;
ii) $M(a)$, defined by (22) in the system with full local memory; and
iii) $T(a)$, Page's CUSUM procedure defined by (9) in the centralized version.

For these three procedures $\tau(a)$, the threshold value a was first determined from the criterion $E_\infty \tau(a) \approx \gamma$. Since $E_\infty N(a)$ is discontinuous (see [2]), the values of γ were chosen so that the corresponding threshold value a exists for each of these procedures. A 10^4-repetition Monte Carlo simulation was performed to determine the appropriate values of a to yield the desired mean time between false alarms γ to within the range of sampling error. Rather than simulating $E_\infty \tau(a)$ for each a separately (which is computationally demanding), an efficient algorithm, suggested by Lorden, is to run *one* simulation to return the record values of the CUSUM statistics and the corresponding values of sample size and then to estimate $E_\infty \tau(a)$ for different a based on these record values.

Next, the renewal property of the CUSUM statistics implies that the detection delay $\overline{E}_1 \tau$ for each of these three procedures is just $E_1 \tau$, the expected sample size when the change happens at time $\nu = 1$. It is therefore straightforward to simulate the detection delay. Monte Carlo experiments with 10^4 repetitions yielded estimates for the detection delays. The results are summarized in Tables I–VI, with the values of a in parentheses.

In the system with two sensors, Tables I and II show that $M(a)$ performs better than $N(a)$ even for moderate γ in both nonsymmetric

TABLE I
TWO NONIDENTICAL SENSORS ($\mu_1 = 0.2$ and $\mu_2 = 1$)

γ	$\overline{E}_1 N(a)$	% DPF	$\overline{E}_1 M(a)$	% DPF	$\overline{E}_1 T(a)$
100	(2.86) 7.7 ± 0.0	31	(2.78) 6.7 ± 0.0	14	(2.87) 5.9 ± 0.0
200	(3.43) 9.6 ± 0.1	35	(3.46) 8.0 ± 0.0	13	(3.52) 7.1 ± 0.0
505	(4.35) 12.2 ± 0.1	39	(4.37) 9.9 ± 0.1	12	(4.41) 8.8 ± 0.1
1000	(5.01) 14.3 ± 0.1	42	(5.05) 11.3 ± 0.1	12	(5.09) 10.1 ± 0.1
2050	(5.66) 16.3 ± 0.1	42	(5.79) 12.8 ± 0.1	11	(5.80) 11.5 ± 0.1
5038	(6.57) 19.0 ± 0.1	44	(6.68) 14.6 ± 0.1	11	(6.68) 13.2 ± 0.1
10046	(7.27) 21.1 ± 0.1	46	(7.36) 16.0 ± 0.1	10	(7.37) 14.5 ± 0.1

(Numbers in parentheses are the values of a so that $E_\infty \tau(a) \approx \gamma$. The decentralized penalty function (DPF) was based on the sampled values.)

TABLE II
TWO IDENTICAL SENSORS ($\mu_1 = \mu_2 = 1$)

γ	$\overline{E}_1 N(a)$	% DPF	$\overline{E}_1 M(a)$	% DPF	$\overline{E}_1 T(a)$
162	(3.34) 5.7 ± 0.0	36	(3.33) 5.4 ± 0.0	29	(3.50) 4.2 ± 0.0
252	(3.86) 6.4 ± 0.0	39	(3.77) 5.9 ± 0.0	28	(3.93) 4.6 ± 0.0
538	(4.68) 7.6 ± 0.0	41	(4.52) 6.8 ± 0.0	26	(4.68) 5.4 ± 0.0
1556	(5.50) 9.2 ± 0.0	44	(5.58) 8.1 ± 0.0	27	(5.73) 6.4 ± 0.0
2154	(5.87) 9.7 ± 0.1	45	(5.90) 8.5 ± 0.0	27	(6.06) 6.7 ± 0.0
5013	(6.76) 11.1 ± 0.1	46	(6.74) 9.5 ± 0.1	25	(6.89) 7.6 ± 0.0
10970	(7.50) 12.2 ± 0.1	47	(7.54) 10.5 ± 0.1	27	(7.69) 8.3 ± 0.0

TABLE III
THREE NONIDENTICAL SENSORS ($\mu_1 = \mu_2 = 0.2$ and $\mu_3 = 1$)

γ	$\overline{E}_1 N(a)$	% DPF	$\overline{E}_1 M(a)$	% DPF	$\overline{E}_1 T(a)$
100	(2.76) 7.6 ± 0.1	31	(2.72) 7.6 ± 0.0	31	(2.89) 5.8 ± 0.0
210	(3.47) 9.6 ± 0.1	35	(3.46) 9.1 ± 0.0	28	(3.60) 7.1 ± 0.0
510	(4.37) 12.0 ± 0.1	38	(4.35) 10.9 ± 0.1	25	(4.45) 8.7 ± 0.0
1018	(4.98) 13.9 ± 0.1	40	(5.05) 12.4 ± 0.1	25	(5.12) 9.9 ± 0.1
2022	(5.70) 15.9 ± 0.1	42	(5.78) 13.9 ± 0.1	24	(5.80) 11.2 ± 0.1
5012	(6.60) 18.5 ± 0.1	44	(6.67) 15.8 ± 0.1	24	(6.70) 12.8 ± 0.1
10076	(7.28) 20.5 ± 0.1	45	(7.37) 17.4 ± 0.1	23	(7.40) 14.1 ± 0.1

TABLE IV
THREE IDENTICAL SENSORS ($\mu_1 = \mu_2 = \mu_3 = 1$)

γ	$\overline{E}_1 N(a)$	% DPF	$\overline{E}_1 M(a)$	% DPF	$\overline{E}_1 T(a)$
160	(3.01) 4.2 ± 0.0	35	(2.94) 4.6 ± 0.0	48	(3.54) 3.1 ± 0.0
1332	(5.50) 6.4 ± 0.0	42	(5.13) 6.6 ± 0.0	47	(5.62) 4.5 ± 0.0
5200	(6.87) 7.8 ± 0.0	44	(6.49) 7.9 ± 0.0	46	(7.00) 5.4 ± 0.0
10600	(7.46) 8.5 ± 0.0	44	(7.21) 8.6 ± 0.0	46	(7.70) 5.9 ± 0.0
20338	(8.24) 9.2 ± 0.0	46	(7.86) 9.2 ± 0.0	46	(8.36) 6.3 ± 0.0
51270	(9.16) 10.2 ± 0.0	48	(8.77) 10.0 ± 0.0	45	(9.27) 6.9 ± 0.0

(Numbers in parentheses are the values of a so that $E_\infty \tau(a) \approx \gamma$. The decentralized penalty function (DPF) was based on the sampled values.)

TABLE V
TEN NONIDENTICAL SENSORS (THREE $\mu_l = 1$AND SEVEN $\mu_l = 0.2$)

γ	$\overline{E}_1 N(a)$	% DPF	$\overline{E}_1 M(a)$	% DPF	$\overline{E}_1 T(a)$
100	(3.16) 3.6 ± 0.0	38	(1.12) 8.4 ± 0.0	223	(3.09) 2.6 ± 0.0
1010	(5.30) 5.7 ± 0.0	43	(3.98) 10.9 ± 0.1	173	(5.33) 4.0 ± 0.0
5112	(6.92) 7.3 ± 0.0	46	(5.72) 12.6 ± 0.1	152	(6.96) 5.0 ± 0.0
10048	(7.59) 7.9 ± 0.0	46	(6.39) 13.3 ± 0.1	144	(7.64) 5.4 ± 0.0
20436	(8.28) 8.6 ± 0.0	48	(7.10) 14.0 ± 0.1	141	(8.35) 5.8 ± 0.0
51046	(9.19) 9.4 ± 0.0	47	(8.04) 14.9 ± 0.1	133	(9.25) 6.4 ± 0.0

TABLE VI
TEN IDENTICAL SENSORS (ALL $\mu_l = 0.2$)

γ	$\overline{E}_1 N(a)$	% DPF	$\overline{E}_1 M(a)$	% DPF	$\overline{E}_1 T(a)$
101	(2.22) 14.6 ± 0.1	29	(1.93) 23.4 ± 0.1	107	(2.48) 11.3 ± 0.1
1003	(4.32) 30.7 ± 0.2	42	(4.66) 50.6 ± 0.2	134	(4.60) 21.6 ± 0.2
5012	(5.91) 43.4 ± 0.2	48	(6.41) 69.2 ± 0.3	136	(6.18) 29.3 ± 0.2
10089	(6.60) 48.7 ± 0.2	49	(7.14) 76.8 ± 0.3	135	(6.86) 32.7 ± 0.2
20193	(7.27) 53.9 ± 0.3	48	(7.86) 84.9 ± 0.4	134	(7.58) 36.3 ± 0.2
50107	(8.19) 61.1 ± 0.3	50	(8.79) 95.2 ± 0.4	133	(8.48) 40.8 ± 0.2

and symmetric systems. In the system with three sensors, Tables III and IV show that for moderate γ, $M(a)$ performs better than $N(a)$ in a nonsymmetric system, but their performances are similar in a symmetric system. In the system with ten sensors, Tables V and VI show that $M(a)$ performs much worse than $N(a)$ for moderate γ in both nonsymmetric and symmetric systems. These are consistent with our asymptotic theory.

It is interesting to see that the DPF function of $M(a)$ seems to be a decreasing function of γ, but the DPF function of $N(a)$ seems to be an increasing function. Comparisons of Tables I–VI indicate that adding sensors with low SNRs actually degrades the performance of $M(a)$ for moderate values of γ, while adding sensors with relatively high SNRs will improve the performance of $M(a)$ for moderate values of γ, but the improvement may not be as good as those of two other procedures $N(a)$ and $T(a)$.

VII. CONCLUSION

We have studied a decentralized extension of quickest change detection problems in two different scenarios. In the system with limited local memory, we have proved the previously conjectured asymptotic optimality of Page's CUSUM procedures with MLRQs under a new condition on observation distributions. The widely used Gaussian or exponential distributions satisfy this condition. In the system with full local memory, we have developed the first of asymptotically optimal procedures. A major theoretical result is that our procedures have same asymptotically first-order performances as the corresponding optimal centralized procedures, although both theoretical analysis and numerical simulations also show that our procedures may perform poorly in some practical situations, especially in the system with large number of sensors, because of the slow asymptotic convergence. It is interesting to note that all these asymptotically optimal decentralized procedures

do not use past messages, and hence neither past message information nor the feedback from the fusion center improves asymptotic performance. Finally, we have compared these asymptotically optimal decentralized procedures with the optimal centralized procedures, especially for Gaussian sensor observations.

There are a number of interesting problems that have not been addressed here. In practice, the distributions of sensor observations often involve unknown parameters. The results developed here are for completely known prechange and postchange distributions, but they provide benchmarks and ideas for the development of procedures in the presence of unknown parameters. It is also of interest to study the system where the observations at the different sensors may be dependent. Moreover, finding fairly simple decentralized procedures that are not only asymptotically optimal but have good performance for practical values of the mean time between false alarms will undoubtedly be of great importance. Therefore, the work in this correspondence should be interpreted as a starting point for further investigation.

APPENDIX
PROOF OF THEOREMS

A. Proof of Theorem 1

In the system with limited local memory, we can rewrite

$$U_{l,n} = \psi_{l,n}(X_{l,n})$$

where $\psi_{l,n}$ may depend on $\boldsymbol{U}_{[1,n-1]} = (U_{1,[1,n-1]}, \ldots, U_{L,[1,n-1]})$. Denote by $f^{\psi}_{l,n}$ and $g^{\psi}_{l,n}$, respectively, the conditional density induced on $U_{l,n}$ given $\boldsymbol{U}_{[1,n-1]}$ when the density of $X_{l,n}$ is f_l and g_l. Denote by $Z_{l,n}$ the conditional log-likelihood ratio function of $U_{l,n}$, $\log\Big(g^{\psi}_{l,n}(U_{l,n})/f^{\psi}_{l,n}(U_{l,n})\Big)$.

Since $X_{1,n}, \ldots, X_{L,n}$ are independent, so are $U_{1,n}, \ldots, U_{L,n}$ given $\boldsymbol{U}_{[1,n-1]}$. Thus, in the fusion center, the conditional log-likelihood ratio of $(U_{1,n}, \ldots, U_{L,n})$ given $\boldsymbol{U}_{1,[1,n-1]}$ is

$$\boldsymbol{Z}_n = \sum_{l=1}^{L} Z_{l,n}.$$

By Theorem 1 of Lai [8], in order to prove (16), it suffices to show that for any $\delta > 0$

$$\lim_{n\to\infty} \sup_{\nu\geq 1} \operatorname{ess\,sup} P_\nu\left\{ \max_{t\leq n} \sum_{k=\nu}^{\nu+t} \boldsymbol{Z}_k \geq I_{\boldsymbol{D}}(1+\delta)n \,\middle|\, \boldsymbol{U}_{[1,\nu-1]} \right\} = 0. \tag{33}$$

By the definition of $I_{D_l}(g_l, f_l)$, for any $k \geq \nu$

$$E_\nu\Big(\boldsymbol{Z}_k \Big| \boldsymbol{U}_{[1,\nu-1]}\Big) = \sum_{l=1}^{L} E_\nu\Big(Z_{l,k}\Big|\boldsymbol{U}_{[1,\nu-1]}\Big) \leq \sum_{l=1}^{L} I_{D_l}(g_l, f_l) = I_{\boldsymbol{D}}$$

and thus

$$\begin{aligned} &P_\nu\left\{ \max_{t\leq n} \sum_{k=\nu}^{\nu+t} \boldsymbol{Z}_k \geq I_{\boldsymbol{D}}(1+\delta)n \,\middle|\, \boldsymbol{U}_{[1,\nu-1]} \right\} \\ &\leq P_\nu\left\{ \max_{t\leq n} \sum_{k=\nu}^{\nu+t} \sum_{l=1}^{L} (Z_{l,k} - E_\nu Z_{l,k}) \geq I_{\boldsymbol{D}}\delta n \,\middle|\, \boldsymbol{U}_{[1,\nu-1]} \right\} \\ &\leq \sum_{l=1}^{L} P_\nu\left\{ \max_{t\leq n} \sum_{k=\nu}^{\nu+t} (Z_{l,k} - E_\nu Z_{l,k}) \geq \delta_1 n \,\middle|\, \boldsymbol{U}_{[1,\nu-1]} \right\} \end{aligned}$$

where $\delta_1 = I_{\boldsymbol{D}}\delta/L$.

Note that $\sum_{k=\nu}^{\nu+t}\Big(Z_{l,k} - E_\nu Z_{l,k}\Big)$ is a martingale under P_ν, Doob's submartingale inequality tells us

$$P_\nu\left\{ \max_{t\leq n} \sum_{k=\nu}^{\nu+t} (Z_{l,k} - E_\nu Z_{l,k}) \geq \delta_1 n \,\middle|\, \boldsymbol{U}_{[1,\nu-1]} \right\} \leq \frac{\sum_{k=\nu}^{\nu+n} E_\nu\Big((Z_{l,k})^2 \,\Big|\, \boldsymbol{U}_{[1,\nu-1]}\Big)}{\delta_1^2 n^2}.$$

By definition, $E_\nu\Big((Z_{l,k})^2 \,\Big|\, \boldsymbol{U}_{[1,\nu-1]}\Big) \leq V_{D_l}(g_l, f_l)$ for any $k \geq \nu$, and hence

$$P_\nu\left\{ \max_{t\leq n} \sum_{k=\nu}^{\nu+t} (Z_{l,k} - E_\nu Z_{l,k}) \geq \delta_1 n \,\middle|\, \boldsymbol{U}_{[1,\nu-1]} \right\} \leq \frac{V_{D_l}(g_l, f_l)}{\delta_1^2 n}$$

which implies (33) since $V_{D_l}(g_l, f_l)$ is finite. Relation (16) follows.

B. Proof of Theorem 2

Assume that $\phi(Y)$ is a quantizer taking values in $\{0, 1\}$. Denote by f_ϕ and g_ϕ, respectively, the density of $\phi(Y)$ when the density of Y is f or g. Let

$$Z_\phi = \log \frac{g_\phi(\phi(Y))}{f_\phi(\phi(Y))}.$$

Note that when $D = 2$

$$E_g(Z_\phi)^2 = \beta_\phi\Big(\log\frac{\beta_\phi}{\alpha_\phi}\Big)^2 + (1-\beta_\phi)\Big(\log\frac{1-\beta_\phi}{1-\alpha_\phi}\Big)^2$$

where $\alpha_\phi = P_f(\phi(Y) = 1)$ and $\beta_\phi = P_g(\phi(Y) = 1)$. Define

$$H(r, s) = r\Big(\log\frac{r}{s}\Big)^2 + (1-r)\Big(\log\frac{1-r}{1-s}\Big)^2$$

for $0 < r, s < 1$ and $H(0,0) = H(1,1) = 0$. To prove $V_2(g, f) < \infty$, it suffices to show that there exists a constant M such that for any ϕ

$$H(\beta_\phi, \alpha_\phi) < M.$$

If one of α_ϕ and β_ϕ is 0 or 1, it is easy to see that Z_ϕ is 0 with probability 1 under g, and hence $H(\beta_\phi, \alpha_\phi) = 0$. Therefore, it suffices to consider the case where $0 < \alpha_\phi, \beta_\phi < 1$. Since $H(b, a) = H(1-b, 1-a)$, assume without loss of generality that $0 < \alpha_\phi \leq \beta_\phi < 1$. (Otherwise consider $1 - \phi(Y)$ and use (18) instead of (17).) Since $1 - B(t)$ is a cumulative distribution function and $B(t)$ is continuous by assumption, there exists $t_0 \in (0, \infty)$ such that

$$B(t_0) = \beta_\phi.$$

Now, let ϕ^* be the likelihood ratio quantizer defined by

$$\phi^* = \begin{cases} 1, & \text{if } g(Y)/f(Y) > t_0 \\ 0, & \text{otherwise.} \end{cases}$$

Then, $P_f(\phi^* = 1) = A(t_0)$ and $P_g(\phi^* = 1) = B(t_0)$.

The proof of Neyman–Pearson lemma [10, p. 65] shows that

$$\int (\phi^* - \phi)\Big(g(y) - t_0 f(y)\Big) d\mu \geq 0$$

so that

$$\Big(B(t_0) - \beta_\phi\Big) - t_0\Big(A(t_0) - \alpha_\phi\Big) \geq 0.$$

Since $B(t_0) = \beta_\phi$ by our choice of t_0, we have

$$A(t_0) \leq \alpha_\phi.$$

Note that for fixed r

$$\frac{\partial H(r,s)}{\partial s} = 2\Big[\frac{1-r}{1-s}\log\frac{1-r}{1-s} - \frac{r}{s}\log\frac{r}{s}\Big]$$

which is positive for all $s \le r$. Thus, $H(r,s)$ is a decreasing function of s in the interval $[0,r]$. In particular

$$H(\beta_\phi, \alpha_\phi) \le H(\beta_\phi, A(t_0)) = H(B(t_0), A(t_0)).$$

Therefore, it suffices to show that there exists a constant M such that for all t

$$H(B(t), A(t)) < M.$$

Since $A(t)$ and $B(t)$ are continuous functions of t, it suffices to show that $H(B(t), A(t))$ is bounded as t goes to 0 or ∞. It is easy to see that if the likelihood ratio $g(y)/f(y)$ has a positive lower bound $C_0 > 0$, then $H(B(t), A(t))$ is 0 if $t < C_0$. Therefore, it suffices to consider the case when such a lower bound does not exist.

Now, $B(t)$ and $A(t)$ go to 1 as t goes to 0, so

$$\lim_{t\to 0}\sqrt{B(t)}\left|\log\frac{B(t)}{A(t)}\right| = 0.$$

By Wald's likelihood ratio identity, we have

$$\begin{aligned} 1 - B(t) &= P_g\left(\frac{g(Y)}{f(Y)} < t\right) = E_f\left(\frac{g(Y)}{f(Y)}; \frac{g(Y)}{f(Y)} < t\right) \\ &\le tP_f\left(\frac{g(Y)}{f(Y)} < t\right) = t(1 - A(t)). \end{aligned}$$

Using the fact that $1 - A(t) \le 1$, we know that $\sqrt{1-B(t)}\left|\log\frac{1-B(t)}{1-A(t)}\right|$ is less than

$$\max\left\{\sqrt{1-B(t)}\left|\log(1-B(t))\right|, \sqrt{1-B(t)}\left|\log t\right|\right\}. \quad (34)$$

As $t \to 0$, $B(t) \to 1$, so that the first term in (34) goes to 0, and by Chebyshev's inequality, the square of the second term is

$$(\log t)^2 P_g\left(\left(-\log\frac{g(Y)}{f(Y)}\right) > |\log t|\right) \le E_g\left(-\log\frac{g(Y)}{f(Y)}\right)^2$$

which is finite by the assumption. Hence,

$$\limsup_{t\to 0} H(B(t), A(t)) < \infty.$$

Similarly, it is clear that

$$\lim_{t\to\infty}\sqrt{1-B(t)}\left|\log\frac{1-B(t)}{1-A(t)}\right| = 0$$

and

$$\limsup_{t\to\infty}\sqrt{B(t)}\left|\log\frac{B(t)}{A(t)}\right| = \limsup_{t\to\infty}\sqrt{B(t)}\left|\log A(t)\right|$$

is finite by the assumption in (17). Hence,

$$\limsup_{t\to\infty} H(B(t), A(t)) < \infty$$

and Theorem 2 is proved.

C. Proof of Theorem 3

To prove (24), define a new stopping time

$$\hat{M}(a) = \inf\left\{n : \sum_{i=1}^{n}\log\frac{g_l(X_{l,i})}{f_l(X_{l,i})} \ge \pi_l a \text{ for all } l = 1, 2, \ldots, L\right\}.$$

By the relation between the one-sided sequential probability ratio tests and Page's CUSUM procedures, it is easy to see that

$$\overline{E}_1 M(a) \le E_1\hat{M}(a) \quad (35)$$

and so it suffices to show that (24) holds for $E_1\hat{M}(a)$. To prove this, for $1 \le l \le L$, let

$$\hat{M}_l = \inf\left\{n : \sum_{i=1}^{n}\log\frac{g_l(X_{l,i})}{f_l(X_{l,i})} \ge \pi_l a\right\}$$

and

$$\tau_l(\hat{M}_l) = \sup\left\{n \ge 1 : \sum_{i=\hat{M}_l+1}^{\hat{M}_l+n}\log\frac{g_l(X_{l,i})}{f_l(X_{l,i})} \le 0\right\}.$$

For simplicity, denote $\tau_l = \tau_l(0)$. It is well known (e.g., in [6, Theorem D]) that for any $1 \le l \le L$

$$E_1\tau_l < \infty \quad (36)$$

since $\log(g_l(X)/f_l(X))$ has positive mean and finite variance under P_1 by Assumption (A2).

By definition of $\hat{M}_l$ and $\tau_l(\hat{M}_l)$, we have

$$\hat{M}(a) \le \max_{1\le l\le L}(\hat{M}_l + \tau_l(\hat{M}_l)) \le \max_{1\le l\le L}\hat{M}_l + \sum_{l=1}^{L}\tau_l(\hat{M}_l).$$

Now since $X_{l,1}, X_{l,2}, \ldots,$ are independent identically distributed (i.i.d.) under P_1, we have $E_1\tau_l(\hat{M}_l) = E_1\tau_l$, and thus

$$E_1\hat{M}(a) \le E_1\max_{1\le l\le L}\hat{M}_l + \sum_{l=1}^{L}E_1\tau_l. \quad (37)$$

By renewal theory and Assumption (A2), under P_1

$$E_1(\hat{M}_l) = \frac{a}{I_{\text{tot}}} + O(1) \text{ and } \text{Var}_1(\hat{M}_l) = O(a)$$

as $a \to \infty$ (see [16] and [17, p. 171]). Hence,

$$\begin{aligned}\left(E_1\left|\hat{M}_l - \frac{a}{I_{\text{tot}}}\right|\right)^2 &\le E_1\left(\hat{M}_l - \frac{a}{I_{\text{tot}}}\right)^2 \\ &= \text{Var}_1(\hat{M}_l) + \left(E_1\hat{M}_l - \frac{a}{I_{\text{tot}}}\right)^2 = O(a)\end{aligned}$$

and so

$$E_1\left|\hat{M}_l - \frac{a}{I_{\text{tot}}}\right| = O\left(\sqrt{a}\right).$$

Thus,

$$\begin{aligned}E_1\max_{1\le l\le L}\hat{M}_l &= \frac{a}{I_{\text{tot}}} + E_1\max_{1\le l\le L}\left(\hat{M}_l - \frac{a}{I_{\text{tot}}}\right) \\ &\le \frac{a}{I_{\text{tot}}} + \sum_{l=1}^{L}E_1\left|\hat{M}_l - \frac{a}{I_{\text{tot}}}\right| = \frac{a}{I_{\text{tot}}} + O(\sqrt{a}).\end{aligned}$$

Relation (24) follows at once from (35)–(37).

To prove (26), let $A = \exp(a)$ and note that

$$\begin{aligned} E_\infty M &= \sum_{n=1}^{\infty} P_\infty(M \ge n) = \sum_{n=1}^{\infty} \int_n^{n+1} P_\infty(M \ge n)dx \\ &\ge \sum_{n=1}^{\infty} \int_n^{n+1} P_\infty(M \ge x)dx = \int_1^{\infty} P_\infty(M \ge x)dx \\ &= A \int_{1/A}^{\infty} P_\infty(M \ge tA)dt. \end{aligned}$$

Thus, by Lemma 6 below and Fatou's lemma

$$\begin{aligned} &\liminf_{a\to\infty}\Big(E_\infty M(a)/A\Big) \\ &\quad \ge \liminf_{a\to\infty} \int_0^{\infty} P_\infty(M(a) \ge tA) 1\left\{t \ge \frac{1}{A}\right\} dt \\ &\quad \ge \int_0^{\infty} \liminf_{a\to\infty} \left[P_\infty(M(a) \ge tA) 1\left\{t \ge \frac{1}{A}\right\}\right] dt \\ &\quad = \int_0^{\infty} \exp(-t)dt = 1 \end{aligned}$$

and hence (26) holds.

To complete the proof, we need to prove the following lemmas.

Lemma 3: Let $W_{l,n}$ be the CUSUM statistic defined in (20). For any l, any $k = 1, 2, \ldots$, and any real number b

$$P_\infty\Big(W_{l,n} \ge b\Big) \le \exp(-b).$$

Proof: For each l, let $S_{l,n}$ denote the log-likelihood ratio $\sum_{i=1}^{n} \log(g_l(X_{l,i})/f_l(X_{l,i}))$, and define $S_{l,0} = 0$. Then, the CUSUM statistic takes the form

$$W_{l,n} = \max_{0\le k<n} (S_{l,n} - S_{l,k}).$$

Since $(X_{l,1}, X_{l,2}, \ldots, X_{l,n})$ have the same joint distribution as $(X_{l,n}, X_{l,n-1}, \ldots, X_{l,1})$, $W_{l,n}$ has the same distribution as $\max_{1\le i\le n} S_{l,i}$. Thus,

$$P_\infty(W_{l,n} \ge b) = P_\infty\Big(\max_{1\le i\le n} S_{l,i} \ge b\Big) = P_\infty(t_l(b) \le n)$$

where

$$t_l(b) = \inf\{n : S_{l,n} \ge b\}.$$

Lemma 3 follows from the fact that

$$P_\infty(t_l(b) \le n) \le P_\infty(t_l(b) < \infty) \le \exp(-b). \qquad \square$$

Lemma 4: For any $k = 1, 2, \ldots$

$$P_\infty(M(a) = k) \le 1/A$$

where $A = \exp(a)$.

Proof: Note that, since the observations are independent from sensor to sensor, application of Lemma 3 yields

$$\begin{aligned} P_\infty(M(a) = k) &\le P_\infty\Big(W_{l,k} \ge \pi_l a \text{ for } 1 \le l \le L\Big) \\ &= \prod_{l=1}^{L} P_\infty(W_{l,k} \ge \pi_l a) \\ &\le \prod_{l=1}^{L} \exp(-\pi_l a) = \exp(-a) = \frac{1}{A}. \end{aligned} \qquad \square$$

Using Lemma 4, it is easy to derive Lemma 5.

Lemma 5: For any $m = 1, 2, \ldots$

$$P_\infty(M(a) \le m) \le \frac{m}{A}.$$

Lemma 6: For $t > 0$

$$\limsup_{a\to\infty} P_\infty(M(a) \le tA) \le 1 - \exp(-t). \tag{38}$$

Proof: For simplicity, we consider only the case when $L = 2$. The same idea can be applied to the cases $L = 1$ and $L \ge 3$. Choose $m = m(a)$ such that $m/a^2 \to \infty$, and $\log m/a \to 0$. Note that

$$\begin{aligned} &P_\infty(M(a) \le tA) \\ &\quad = P_\infty\left(\max_{0\le k<tA/m} \max_{km+1\le j\le(k+1)m} \left[\min_{1\le l\le 2} \frac{W_{l,j}}{\pi_l}\right] > a\right) \\ &\quad = P_\infty\left(\max_k \max_j \left[\min_{1\le l\le 2} \max_{i_l} \frac{S_{l,j} - S_{l,i_l}}{\pi_l}\right] > a\right) \end{aligned} \tag{39}$$

where the maximum is taken over $0 \le k < tA/m$, $km + 1 \le j \le (k+1)m$, and $1 \le i_l \le j$ for all $l = 1, 2$. For all such k, define

$$\begin{aligned} C_1(k) &= \{i_1 : km + 1 \le i_1 \le j \le (k+1)m\} \\ C_2(k) &= \{i_1 : 1 \le i_1 \le km\} \\ D_1(k) &= \{i_2 : km + 1 \le i_2 \le j \le (k+1)m\} \\ D_2(k) &= \{i_2 : 1 \le i_2 \le km\}. \end{aligned}$$

For simplicity, omit k, e.g., write C_1 for $C_1(k)$, and define

$$\begin{aligned} B_1 &= C_1 \cap D_1, \quad B_2 = C_2 \cap D_1, \\ B_3 &= C_1 \cap D_2, \quad B_4 = C_2 \cap D_2. \end{aligned}$$

For $r = 1, 2, 3, 4$, denote

$$Q_r = P_\infty\left(\max_k \max_j \left[\min_{1\le l\le 2} \max_{B_r} \frac{S_{l,j} - S_{l,i_l}}{\pi_l}\right] > a\right)$$

where the maximum is taken over $0 \le k < tA/m$, $km + 1 \le j \le (k+1)m$, and $(i_1, i_2) \in B_r$. Note that the right-hand side of (39) is less than $\sum_{r=1}^{4} Q_r$, and hence it suffices to show that

$$\limsup_{a\to\infty} \sum_{r=1}^{4} Q_r \le 1 - \exp(-t).$$

It is easy to see that

$$Q_1 = 1 - \prod_k P_\infty\left(\max_j \left[\min_{1\le l\le 2} \max_{i_l} \frac{S_{l,j} - S_{l,i_l}}{\pi_l}\right] \le a\right)$$

where the product is taken over $0 \le k < tA/m$, and the maximum is taken over $km + 1 \le i_l \le j \le (k+1)m$ for all $l = 1, 2$. Thus,

$$Q_1 = 1 - \Big(P_\infty(M(a) > m)\Big)^{tA/m}.$$

By Lemma 5, we have

$$Q_1 \le 1 - \left(1 - \frac{m}{A}\right)^{tA/m}.$$

Note that since $m/A \to 0$ as $a \to \infty$, for given $\delta > 0$, once a is sufficiently large

$$1 - \frac{m}{A} \ge \exp\left(-(1+\delta)\frac{m}{A}\right)$$

and thus $Q_1 \le 1 - \exp(-(1+\delta)t)$. Letting $\delta \to 0$, we obtain

$$\limsup_{a\to\infty} Q_1 \le 1 - \exp(-t).$$

To complete the proof of Lemma 6, it suffices to show that for all $\epsilon > 0$, Q_2, Q_3 and Q_4 are smaller than ϵ for sufficiently large a. We will prove this fact for Q_2 in Lemma 7. The proofs for Q_3 and Q_4 are similar. □

Lemma 7: Under the condition (25) of Theorem 3, for all $\epsilon > 0$, once a is sufficiently large

$$Q_2 = P_\infty\left(\max_k \max_j \left[\min_{1\le l\le 2} \max_{i_l} \frac{S_{l,j} - S_{l,i_l}}{\pi_l}\right] > a\right) \le \epsilon$$

where the maximum is taken over $0 \le k < tA/m$, $km+1 \le j \le (k+1)m$, $1 \le i_1 \le km$, and $km+1 \le i_2 \le j \le (k+1)m$.

Proof: Note that $j - i_1 = j - km + km - i_1$ and $S_{1,j} - S_{1,i_1}$ equals to the sum of the independent random walks $S_{1,j} - S_{1,km}$ and $S_{1,km} - S_{1,i_1}$. Hence, if $\{\overline{S}_i\}$ is an independent copy of $\{S_{1,i}\}$, then

$$\begin{aligned} Q_2 &\le \frac{tA}{m}\sum_{j=1}^{m} P_\infty\left(\max_{0\le i\le tA} \overline{S}_i + S_{1,j} > \pi_1 a \text{ and } W_{2,j} > \pi_2 a\right) \\ &\le \frac{tA}{m}\sum_{j=1}^{m} P_\infty\left(\max_{0\le i\le tA} \overline{S}_i + S_{1,j} > \pi_1 a\right) P_\infty(W_{2,j} > \pi_2 a) \\ &\le \frac{t\exp(\pi_1 a)}{m}\sum_{j=1}^{m} P_\infty\left(\max_{0\le i\le tA} \overline{S}_i + S_{1,j} > \pi_1 a\right) \end{aligned}$$

using Lemma 3 for $W_{2,j}$.

Now, using Wald's likelihood ratio identity

$$\begin{aligned} &P_\infty\left(\max_{0\le i\le tA} \overline{S}_i + S_{1,j} > \pi_1 a\right) \\ &\le P_\infty(S_{1,j} > \pi_1 a) + P_\infty\left(\max_{0\le i\le tA} \overline{S}_i > \pi_1 a - S_{1,j} > 0\right) \\ &\le P_\infty(S_{1,j} > \pi_1 a) + E_\infty\Big(\exp(S_{1,j} - \pi_1 a); \pi_1 a - S_{1,j} > 0\Big) \\ &= E_\infty \exp\Big(\min(0, S_{1,j} - \pi_1 a)\Big) \\ &= E_1\Big[\exp(-S_{1,j}) \cdot \exp\Big(\min(0, S_{1,j} - \pi_1 a)\Big)\Big]. \end{aligned}$$

Thus,

$$Q_2 \le \frac{t}{m}\sum_{j=1}^{m} E_1 \exp\Big(\min(\pi_1 a - S_{1,j}, 0)\Big).$$

Applying Lemma 8 (below) for $S_{1,j}$ under P_1, and letting $m_1 = a^2$, we have for sufficiently large a

$$\sup_{j\ge m_1} E_1 \exp\Big(\min(\pi_1 a - S_{1,j}, 0)\Big) \le \epsilon_1$$

Therefore,

$$Q_2 \le \frac{t}{m}\Big(m_1 \cdot 1 + (m - m_1)\epsilon_1\Big) \le t\Big(\frac{m_1}{m} + \epsilon_1\Big)$$

and the lemma follows, since the right-hand side goes to 0 as a goes to ∞. □

Lemma 8: Suppose $X_1, X_2, \ldots$ are i.i.d. with $EX_i = \mu > 0$, $\mathrm{Var}(X_i) = \sigma^2$, and $E|X_i|^3 = \rho < \infty$. Let $S_n = X_1 + \cdots + X_n$ and $m_1 = b^2$. Then

$$\sup_{n\ge m_1} E\exp\Big((\min(b - S_n, 0)\Big) \to 0$$

as $b \to \infty$.

Proof: First, we establish

$$\begin{aligned} E\exp\Big((\min(b - S_n, 0)\Big) &\le \frac{3\rho}{\sigma^3\sqrt{n}} + \Phi\left(\frac{b - n\mu}{\sigma\sqrt{n}}\right) \\ &\quad + A\left(\frac{b - n\mu}{\sigma\sqrt{n}} + \sigma\sqrt{n}\right)\exp\left(b + (\frac{\sigma^2}{2} - \mu)n\right) \end{aligned} \quad (40)$$

where $\Phi(x)$ is the standard Gaussian distribution and $A(x) = \Phi(-x) = 1 - \Phi(x)$.

Let $F_n(x)$ denote the distribution function of S_n, then

$$\left|F_n(x) - \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)\right| \le \frac{3\rho}{\sigma^3\sqrt{n}}$$

for any x by the Berry–Esseen theorem. Now

$$\begin{aligned} &E\exp\Big((\min(b - S_n, 0)\Big) \\ &= F_n(b) + \int_b^\infty \exp(b - x)dF_n(x) \\ &= \int_b^\infty F_n(x)\exp(b - x)dx \\ &\le \int_b^\infty \left(\frac{3\rho}{\sigma^3\sqrt{n}} + \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)\right)\exp(b - x)dx \\ &= \frac{3\rho}{\sigma^3\sqrt{n}} + \Phi\left(\frac{b - n\mu}{\sigma\sqrt{n}}\right) + \int_b^\infty \phi\Big(\frac{x - n\mu}{\sigma\sqrt{n}}\Big)\exp(b - x)\frac{1}{\sigma\sqrt{n}}dx \end{aligned}$$

and hence (40) holds.

We next bound each term on the right-hand side of (40). For $n \ge m_1$, the first two terms are uniformly bounded by

$$\frac{3\rho}{\sigma^3 b} + \Phi\left(\frac{1 - \mu b}{\sigma}\right)$$

which goes to 0 as $b \to \infty$.

For the third term on the right-hand side of (40), we need to consider two cases: 1) $\mu > \sigma^2/2$ and 2) $\mu \le \sigma^2/2$. In case 1), note that $A(x) \le 1$, and so for all $n \ge m_1$, the third term is smaller than

$$\exp\left(b - \left(\mu - \frac{\sigma^2}{2}\right)b^2\right)$$

which goes to 0 as $b \to \infty$. In case 2), note that $A(x) \le \phi(x)/x$ for all $x > 0$, where $\phi(x)$ is the density function of the standard Gaussian distribution (see [26, p. 141]). Thus, the third term is smaller than

$$\frac{\sigma\sqrt{n}}{b + (\sigma^2 - \mu)n}\phi\left(\frac{b - \mu n}{\sigma\sqrt{n}}\right)$$

which also goes to 0 uniformly for all $n \ge m_1$ as $b \to \infty$.

Therefore, Lemma 8 holds. □

ACKNOWLEDGMENT

The author would like to thank his advisor Dr. Gary Lorden, for his constant support and encouragement, Dr. Venugopal V. Veeravalli for bringing this problem to his attention, as well as Dr. Alexander G. Tartakovsky for fruitful discussions. The author also would like to thank the referees for helpful suggestions, which led to significant improvements in organization and presentation.

REFERENCES

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

[2] R. W. Crow and S. C. Schwartz, "Quickest detection for sequential decentralized decision systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 1, pp. 267–283, Jan. 1996.

[3] V. P. Dragalin, "Asymptotics for a sequential selection procedure," *Statist. Decisions*, pp. 123–137, 1999. Suppl. Issue no. 4.

[4] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests—Part II: Accurate asymptotic expansions for the expected sample size," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1366–1383, Jul. 2000.

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: Wiley, 1971, vol. II.

[6] J. Kiefer and J. Sacks, "Asymptotically optimal sequential inference and design," *Ann. Math. Statist.*, vol. 34, pp. 705–750, 1963.

[7] T. L. Lai, "Sequential change-point detection in quality control and dynamical systems," *J. Roy. Statist. Soc. Ser. B*, vol. 57, pp. 613–658, 1995.

[8] ——, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2917–2929, Nov. 1998.
[9] ——, "Sequential multiple hypothesis testing and efficient fault detection-isloation in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 595–608, Mar. 2000.
[10] E. L. Lehmann, *Testing Statistical Hypothesis*. New York: Wiley, 1959.
[11] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, pp. 1897–1908, 1971.
[12] Y. Mei, "Asymptotically optimal methods for sequential change-point detection," Ph.D. dissertation, Calif. Inst. Technol., Pasadena, CA, 2003.
[13] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, pp. 1379–1387, 1986.
[14] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.
[15] M. Pollak, "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 13, pp. 206–227, 1985.
[16] D. Siegmund, "The variance of one-sided stopping rules," *Ann. Math. Statist.*, vol. 40, pp. 1074–1077, 1969.
[17] ——, *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer-Verlag, 1985.
[18] D. Siegmund and E. S. Venkatraman, "Using the generalized likelihood ratio statistics for sequential detection of a change-point," *Ann. Statist.*, vol. 23, pp. 255–271, 1995.
[19] A. N. Shiryayev, "On optimum methods in quickest detection problems," *Theory Probab. Appl.*, vol. 8, pp. 22–46, 1963.
[20] ——, *Optimal Stopping Rules*. New York: Springer-Verlag, 1978.
[21] A. G. Tartakovsky and V. V. Veeravalli, "An efficient sequential procedure for detecting changes in multichannel and distributed systems," in *Proc. 5th Int. Conf. Information Fusion*, vol. 2, Annapolis, MD, Jul. 2002, pp. 1–8.
[22] J. N. Tsitsiklis, "Extremal properties of likelihood ratio quantizers," *IEEE Trans. Commun.*, vol. 41, no. 4, pp. 550–558, Apr. 1993.
[23] V. V. Veeravalli, "Sequential decision fusion: Theory and applications," *J. Franklin Inst.*, vol. 336, pp. 301–322, Feb. 1999.
[24] ——, "Decentralized quickest change detection," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1657–1665, May 2001.
[25] V. V. Veeravalli, T. Basar, and H. V. Poor, "Decentralized sequential detection with a fusion center performing the sequential test," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 433–442, Mar. 1993.
[26] D. Williams, *Probabilty with Martingles*. Cambridge, U.K.: Cambridge Univ. press, 1991.

Efficient Computation of the Hidden Markov Model Entropy for a Given Observation Sequence

Diego Hernando, Valentino Crespi, *Member, IEEE*, and George Cybenko, *Fellow, IEEE*

***Abstract*—Hidden Markov models (HMMs) are currently employed in a wide variety of applications, including speech recognition, target tracking, and protein sequence analysis. The Viterbi algorithm is perhaps the best known method for tracking the hidden states of a process from a sequence of observations. An important problem when tracking a process with an HMM is estimating the uncertainty present in the solution. In this correspondence, an algorithm for computing at runtime the entropy of the possible hidden state sequences that may have produced a certain sequence of observations is introduced. The brute-force computation of this quantity requires a number of calculations exponential in the length of the observation sequence. This algorithm, however, is based on a trellis structure resembling that of the Viterbi algorithm, and permits the efficient computation of the entropy with a complexity linear in the number of observations.**

***Index Terms*—Entropy, hidden Markov model (HMM), performance measurement, process query system, Viterbi algorithm.**

I. Introduction

Hidden Markov models (HMMs) are often used to find the most likely hidden state sequence that produces a given sequence of observations. This can be done with the well-known Viterbi algorithm. Possible performance measures in this scenario include the probability of error on a single state and the probability of error on the whole sequence. An alternative measure is the entropy of the possible *solutions* (state sequences) that explain a certain observation sequence.

The entropy of a random variable provides a measure of its uncertainty. The entropy of the state sequence that explains an observation sequence, given a model, can be viewed as the minimum number of bits that, on average, will be needed to encode the state sequence (given the model and the observations) [1]. The higher this entropy, the higher the uncertainty involved in tracking the hidden process with the current model.

In this correspondence, we introduce an efficient algorithm for computing at runtime the entropy of the hidden state sequence that explains a given observation sequence.

The remainder of this document is organized as follows: Section II gives a brief introduction to HMMs and specifies the notation used in this document. Section III describes the algorithm for efficiently computing the entropy at runtime, along with a numerical example and a brief analysis of the algorithm's performance, in terms of the number of operations required. Finally, Section IV contains the conclusions and a discussion of the usefulness of our algorithm.

Manuscript received February 18, 2004; revised March 27, 2005. This work was supported in part by ARDA under Grant F30602-03-C-0248, DARPA Projects F30602-00-2-0585 and F30602-98-2-0107, and the National Institute of Justice, Department of Justice Award number 2000-DT-CX-K001.

D. Hernando is with the Electrical and Computer Engineering Department, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: diego.hernando@ieee.org).

V. Crespi is with the Department of Computer Science at the California State University, Los Angeles, CA 90032-8150 USA (e-mail: vcrespi@calstatela.edu).

G. Cybenko is with the Thayer School of Engineering, Dartmouth College, Hanover, NH 03755 USA (e-mail: george.cybenko@dartmouth.edu).

Communicated by X. Wang, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2005.850223