Tracking Stopping Times Through Noisy Observations

Urs Niesen, *Student Member, IEEE*, and Aslan Tchamkerten

***Abstract*—A novel quickest detection setting is proposed, generalizing the well-known Bayesian change-point detection model. Suppose $\{(X_i, Y_i)\}_{i\geq 1}$ is a sequence of pairs of random variables, and that S is a stopping time with respect to $\{X_i\}_{i\geq 1}$. The problem is to find a stopping time T with respect to $\{Y_i\}_{i\geq 1}$ that optimally tracks S, in the sense that T minimizes the expected *reaction delay* $\mathbb{E}(T-S)^+$, while keeping the *false-alarm probability* $\mathbb{P}(T<S)$ below a given threshold $\alpha \in [0,1]$. This problem formulation applies in several areas, such as in communication, detection, forecasting, and quality control.**

Our results relate to the situation where the X_i's and Y_i's take values in finite alphabets and where S is bounded by some positive integer κ. By using elementary methods based on the analysis of the tree structure of stopping times, we exhibit an algorithm that computes the optimal average reaction delays for all $\alpha \in [0,1]$, and constructs the associated optimal stopping times T. Under certain conditions on $\{(X_i, Y_i)\}_{i\geq 1}$ and S, the algorithm running time is polynomial in κ.

***Index Terms*—Algorithms, decision theory, feedback communication, forecasting, monitoring, optimal stopping, quickest detection problem, sequential analysis, synchronization, tree analysis.**

I. Problem Statement

THE tracking stopping time (TST) problem is defined as follows. Let $\{(X_i, Y_i)\}_{i\geq 1}$ be a sequence of pairs of random variables. Alice observes $X_1, X_2, \ldots$ and chooses a stopping time (s.t.) S with respect to that sequence.[1] Knowing the distribution of $\{(X_i, Y_i)\}_{i\geq 1}$ and the stopping rule S, but having access only to the Y_i's, Bob wishes to find a s.t. that gets as close as possible to Alice's. Specifically, Bob aims to find a s.t. T minimizing the expected reaction delay $\mathbb{E}(T-S)^+ \triangleq \mathbb{E}\max\{0, T-S\}$, while keeping the false-alarm probability $\mathbb{P}(T<S)$ below a certain threshold $\alpha \in [0,1]$.

Manuscript received February 06, 2008. Current version published December 24, 2008. This work was supported in part by the NSF under Grant CCF-0515122, by DoD MURI under Grant N00014-07-1-0738, and by a University IR&D Grant from Draper Laboratory. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Nice, France, June 2007. This work was carried out while A. Tchamkerten was at MIT as a Postdoctoral Associate.

U. Niesen is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139 USA (e-mail: uniesen@mit.edu).

A. Tchamkerten is with the Department of Communications and Electronics, TELECOM ParisTech, 75634 Paris Cedex 13, France (e-mail: tcham@mit.edu).

Communicated by H. Bölcskei, Associate Editor for Detection and Estimation.

Digital Object Identifier 10.1109/TIT.2008.2008115

[1]Recall that a s.t. with respect to a sequence of random variables $\{X_i\}_{i\geq 1}$ is a random variable S taking values in the positive integers such that the event $\{S=n\}$, conditioned on $\{X_i\}_{i=1}^n$, is independent of $\{X_i\}_{i=n+1}^\infty$ for all $n \geq 1$. An s.t. S is *nonrandomized* if $\mathbb{P}(S=n|X^n=x^n) \in \{0,1\}$ for all $x^n \in \mathcal{Y}^n$ and $n \geq 1$. An s.t. S is *randomized* if $\mathbb{P}(S=n|X^n=x^n) \in [0,1]$ for all $x^n \in \mathcal{X}^n$ and $n \geq 1$.

Example 1. Monitoring: Let X_i be the distance of an object from a barrier at time i, and let S be the first time the object hits the barrier, i.e., $S \triangleq \inf\{i \geq 1 : X_i = 0\}$. Assume we have access to X_i only through a noisy measurement Y_i, and that we want to raise an alarm as soon as the object hits the barrier. This problem can be formulated as the one of finding a s.t. T with respect to the Y_i's that minimizes the expected reaction delay $\mathbb{E}(T-S)^+$, while keeping the false-alarm probability $\mathbb{P}(T<S)$ small enough. ◇

Another situation where the TST problem applies is in the context of communication over channels with feedback. Most of the studies related to feedback communication assume perfect feedback, i.e., the transmitter is fully aware of the output of the channel as observed by the receiver. Without this assumption—i.e., if the feedback link is noisy—a synchronization problem may arise between the transmitter and the receiver which can be formulated as a TST problem, as shown in the following example.

Example 2. Communication: It is well known that the presence of a noiseless feedback link allows to dramatically increase the reliability for a given communication delay (see, e.g., [12]). However, to take advantage of feedback, variable-length codes are often necessary.[2] This can be observed by looking at a nonperfect binary erasure channel. In this case, any block coding strategy yields a strictly positive error probability. In contrast, consider the variable-length strategy where the encoder keeps sending the bit it wishes to convey until it is successfully received. This simple strategy achieves error-free communication at a rate equal to the capacity of the channel in question. Can we still use this coding strategy if the feedback channel is (somewhat) noisy? Because of the noisy feedback link, a synchronization problem between the decoder and the encoder arises: if the first nonerased output symbol occurs at time S, what should be sent at time $S+1$? This agreement problem occurs because the encoder observes now only a noisy version of the symbols received by the decoder (see Fig. 1). In particular, the first nonerased output symbol may not be recognized as such by the encoder.[3] Instead of treating the synchronization issue that results from the use of variable-length codes over channels with noisy feedback, let us consider the simpler problem of finding the minimum delay needed by the encoder to realize that the de-

[2]The reliability function associated with block coding schemes is lower than the one associated with variable length coding. For symmetric channels, for instance, the reliability function associated with block coding schemes is limited by the sphere packing bound, which is lower than the best optimal error exponent attainable with variable length coding ([5], [10]).

[3]For fixed-length coding strategies over channels with noisy feedback we refer the reader to [13], [6].

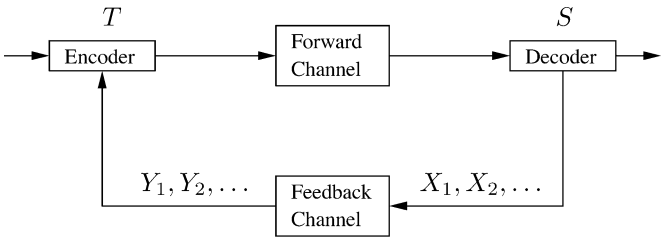

Fig. 1. The decoding time S depends on the output of the forward channel. The encoder decides to stop transmission at time T based on the output of the feedback channel. If the feedback channel is noisy, S and T need not coincide.

coder has made a decision. In terms of the TST problem, Alice and Bob represent the decoder and the encoder, the X_i's and Y_i's correspond to the input and output symbols of the feedback channel, whereas S and T represent the decoding time and the time the encoder stops transmission, respectively. Here $\mathbb{E}(T-S)^+$ represents the delay needed by the encoder to realize that the decoder has made a decision, and we aim to minimize it given that the probability of stopping transmission too early, $\mathbb{P}(T < S)$, is kept below a certain threshold α.

Note that , in the context of feedback communication, it is reasonable to define the communication rate with respect to the overall delay $S + (T - S)^+ = \max\{S, T\}$. This definition, in contrast with the one that takes into account only the decoding time (such as for rateless codes), puts the delay constraint on both the transmitter and the receiver. In Example 8, we investigate the highest achievable rate with respect to the overall communication delay if the "send until a non-erasure occurs" strategy is used and both the forward and the feedback links are binary erasure channels. ◇

Example 3. Forecasting: A large manufacturing machine breaks down as soon as its cumulative fatigue hits a certain threshold. Knowing that a machine replacement takes, say, ten days, the objective is to order a new machine so that it is operational at the time the old machine breaks down. This prevents losses due to an interrupted manufacturing process as well as storage costs caused by an unused backup machine.

The problem of determining the operating start date of the new machine can be formulated as follows. Let X_n be the cumulative fatigue up to day n of the current machine, and let S denote the first day n that X_n crosses the critical fatigue threshold. Since the replacement period is ten days, the first day T a new machine is operational can be scheduled only on the basis of a (possibly randomized) function of $\{X_i\}_{i=1}^{T-10}$. By defining Y_i to be equal to X_{i-10} if $i > 10$ and else equal to zero, the day T is now a s.t. with respect to $\{Y_i\}_{i\geq 1}$, and we can formulate the requirement on T as aiming to minimize $\mathbb{E}(T - S)^+$ while keeping $\mathbb{P}(T < S)$ below a certain threshold. ◇

In the forecasting example, Alice has access to more information than Bob. From the process she observes, she can deduce Bob's observations—simply by delaying hers. This feature may be interesting in other applications. The general formulation where Alice has access to more information than Bob is obtained by letting the observation available to Alice at time i be $X_i = (\widetilde{X}_i, \widetilde{Y}_i)$, and the observation available to Bob be $Y_i = \widetilde{Y}_i$.

Example 4. Bayesian Change-Point Detection: In this example, we will see how the TST setting generalizes the Bayesian version of the change-point detection problem, a long studied problem with applications to industrial quality control and that dates back to the 1940s [1]. The Bayesian change-point problem is formulated as follows. Let θ be a random variable taking values in the positive integers. Let $\{Y_i\}_{i\geq 1}$ be a sequence of random variables such that, given the value of θ, the conditional probability of Y_n given $Y^{n-1} \triangleq \{Y_i\}_{i=1}^{n-1}$ is $P_0(\cdot|Y^{n-1})$ for $n < \theta$ and is $P_1(\cdot|Y^{n-1})$ for $n \geq \theta$. We are interested in a s.t. T with respect to the Y_i's minimizing the change-point reaction delay $\mathbb{E}(T-\theta)^+$, while keeping the false-alarm probability $\mathbb{P}(T < \theta)$ below a certain threshold $\alpha \in [0, 1]$.

Shiryaev (see, e.g.,[20], [19, Ch. 4.3]) considered the Lagrangian formulation of the above problem: Given a constant $\lambda \geq 0$, minimize

$$\mathbb{E}(T - \theta)^+ + \lambda \mathbb{P}(T < \theta)$$

over all s.t.'s T. Assuming a geometric prior on the change-point θ and that before and after θ the observations are independent with common density function f_0 for $t < \theta$ and f_1 for $t \geq \theta$, Shiryaev showed that the optimal T stops as soon as the posterior probability that a change occurred exceeds a certain fixed threshold. Later Yakir [22] generalized Shiryaev's result by considering finite-state Markov chains. For more general prior distributions on θ, the problem is known to become difficult to handle. However, in the limit $\alpha \to 0$, Lai [14] and, later, Tartakovsky and Veeravalli [21], derived asymptotically optimal detection policies for the Bayesian change-point problem under general assumptions on the distributions of the change-point and observed process.[4]

To see that the Bayesian change-point problem can be formulated as a TST problem, it suffices to define the sequence of binary random variables $\{X_i\}_{i\geq 1}$ such that $X_i = 0$ if $i < \theta$ and $X_i = 1$ if $i \geq \theta$, and to let $S \triangleq \inf\{i : X_i = 1\}$ (i.e., $S = \theta$). The change-point problem defined by θ and $\{Y_i\}_{i\geq 1}$ becomes the TST problem defined by S and $\{(X_i, Y_i)\}_{i\geq 1}$. However, the TST problem cannot, in general, be formulated as a Bayesian change-point problem. Indeed, the Bayesian change-point problem yields for any $k > n$

$$\begin{aligned}
&\mathbb{P}(\theta = k|Y^n = y^n, \theta > n) \\
&\quad = \frac{\mathbb{P}(Y^n = y^n, \theta > n|\theta = k)\mathbb{P}(\theta = k)}{\mathbb{P}(Y^n = y^n|\theta > n)\mathbb{P}(\theta > n)} \\
&\quad = \frac{\mathbb{P}(Y^n = y^n|\theta = k)\mathbb{P}(\theta = k)}{\mathbb{P}(Y^n = y^n|\theta > n)\mathbb{P}(\theta > n)} \\
&\quad = \mathbb{P}(\theta = k|\theta > n) \qquad (1)
\end{aligned}$$

since $\mathbb{P}(Y^n = y^n|\theta = k) = \mathbb{P}(Y^n = y^n|\theta > n)$. In words, given that no change occurred up to time n, the observations y^n are useless in predicting the value of the change-point θ. In contrast, for the TST problem, in general we have

$$\mathbb{P}(S = k|Y^n = y^n, S > n) \neq \mathbb{P}(S = k|S > n) \qquad (2)$$

because $\mathbb{P}(Y^n = y^n|S = k) \neq \mathbb{P}(Y^n = y^n|S > n)$. ◇

[4]For the non-Bayesian version of the change-point problem, we refer the reader to [16], [18], [17].

As is argued in the last example, the TST problem is a generalization of the Bayesian change-point problem, which itself is analytically tractable only in special cases. This makes an analytical treatment of the general TST problem difficult. Instead, we present an algorithmic solution to this problem for an arbitrary process $\{(X_i,Y_i)\}_{i\geq 1}$ and an arbitrary s.t. S bounded by some constant $\kappa \geq 1$. The proof of correctness of this algorithm provides insights into the structure of the optimal s.t. T tracking S, and into the tradeoff between expected delay $\mathbb{E}(T-S)^+$ and probability of false-alarm $\mathbb{P}(T<S)$. Under some conditions on $\{(X_i,Y_i)\}_{i\geq 1}$ and S, the computational complexity of this algorithm is polynomial in κ.

The rest of the paper is organized as follows. In Section II, we provide some basic properties of the TST problem defined over a finite alphabet process $\{(X_i,Y_i)\}_{i\geq 1}$, and in Section III, we provide an algorithmic solution to it. In Section IV, we derive conditions under which the algorithm has low complexity and illustrate this in Section V with examples.

II. The Optimization Problem

Let $\{(X_i,Y_i)\}_{i\geq 1}$ be a discrete-time process where the X_i's and Y_i's take value in some finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let S be a s.t. with respect to $\{X_i\}_{i\geq 1}$ such that $S\leq\kappa$ almost surely for some constant $\kappa\geq 1$. We aim to find for any $\alpha\in[0,1]$

$$d(\alpha)\triangleq \min_{\substack{T:\mathbb{P}(T<S)\leq\alpha\\ T\leq\kappa}} \mathbb{E}(T-S)^+ \tag{3}$$

where the s.t.'s T are possibly randomized. Note that the restriction $T\leq\kappa$ is without loss of optimality.

Now, the set of all s.t.'s over $\{Y_i\}_{i\geq 1}$ is convex, and its extreme points are nonrandomized s.t.'s ([2], [11]). This implies that any randomized s.t. $T\leq\kappa$ can be written as a convex combination of nonrandomized s.t.'s bounded by κ, i.e.,

$$\mathbb{P}(T=k)=\sum_j w_j\mathbb{P}(T_j=k)$$

for any integer k, where $\{T_j\}$ denotes the finite set of all nonrandomized s.t.'s bounded by κ, and where the w_j's are nonnegative and sum to one. Hence, because false-alarm and expected reaction delay can be written as

$$\mathbb{P}(T<S)=\sum_j w_j\mathbb{P}(T_j<S)$$

$$\mathbb{E}(T-S)^+=\sum_j w_j\mathbb{E}(T_j-S)^+$$

the function $d(\alpha)$ is convex and piecewise linear, with breakpoints achieved by nonrandomized s.t.'s. Its typical shape is depicted in Fig. 2.

For $\lambda\geq 0$, define the Lagrangian

$$J_\lambda(T)\triangleq\mathbb{E}(T-S)^+ +\lambda\mathbb{P}(T<S). \tag{4}$$

Lemma 1: We have

$$d(\alpha)=\sup_{\lambda\geq 0}\min_{T\leq\kappa}(J_\lambda(T)-\lambda\alpha)$$

where the minimization is over all nonrandomized s.t.'s bounded by κ.

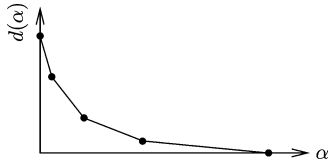

Fig. 2. Typical shape of the expected delay $d(\alpha)$ as a function of false-alarm probability α. The breakpoints are achieved by nonrandomized s.t.'s.

Proof: The linear (in $\{w_j\}$) minimization problem (3) admits at least one feasible point, namely, $T=\kappa$. Therefore, strong Lagrange duality holds (see, e.g., [3, Ch. 5]), and we obtain

$$d(\alpha)=\sup_{\lambda\geq 0}\min_{T\leq\kappa}(J_\lambda(T)-\lambda\alpha). \tag{5}$$

Because $d(\alpha)$ is convex with extreme points achieved by nonrandomized s.t.'s, we may restrict the minimization in (5) to be over the set of nonrandomized s.t.'s bounded by κ. □

III. An Algorithm For Computing $d(\alpha)$

We first establish a few preliminary results later used to evaluate $\min_T J_\lambda(T)$. Emphasis is put on the finite tree representation of bounded s.t.'s with respect to finite alphabet processes. We then provide an algorithm that computes the entire curve $d(\alpha)$.

We introduce a few notational conventions. The set $\mathcal{Y}^*$ represents all finite sequences over $\mathcal{Y}$. An element in $\mathcal{Y}^*$ is denoted either by y^n or by $\boldsymbol{y}$, depending on whether or not we want to emphasize its length. To any nonrandomized s.t. T, we associate a unique $|\mathcal{Y}|$-ary tree $\mathcal{T}$ (i.e., all the nodes of $\mathcal{T}$ have either zero or exactly $|\mathcal{Y}|$ children) having each node specified by some $\boldsymbol{y}\in\mathcal{Y}^*$, where $\rho\boldsymbol{y}$ represents the vertex path from the root ρ to the node $\boldsymbol{y}$. The depth of a node $y^n\in\mathcal{T}$ is denoted by $l(y^n)\triangleq n$. The tree consisting only of the root is the trivial tree. A node $y^n\in\mathcal{T}$ is a leaf if $\mathbb{P}(T=n|Y^n=y^n)=1$. We denote by $\mathcal{L}(\mathcal{T})$ the leaves of $\mathcal{T}$ and by $\mathcal{I}(\mathcal{T})$ the intermediate (or nonterminal) nodes of $\mathcal{T}$. The notation $T(\mathcal{T})$ is used to denote the (nonrandomized) s.t. T induced by the tree $\mathcal{T}$. Given a node $\boldsymbol{y}$ in $\mathcal{T}$, let $\mathcal{T}_{\boldsymbol{y}}$ be the subtree of $\mathcal{T}$ rooted in $\boldsymbol{y}$. Finally, let $\mathcal{D}(\mathcal{T}_{\boldsymbol{y}})$ denote the descendants of $\boldsymbol{y}$ in $\mathcal{T}$. The next example illustrates these notations.

Remark: Unless explicitly stated otherwise, trees and their subtrees have the same root.

Example 5: Let $\mathcal{Y}=\{0,1\}$ and $\kappa=2$. The tree $\mathcal{T}$ depicted in Fig. 3 corresponds to the nonrandomized s.t. T taking value 1 if $Y_1=1$ and value 2 if $Y_1=0$. The sets $\mathcal{L}(\mathcal{T})$ and $\mathcal{I}(\mathcal{T})$ are given by $\{00,01,1\}$ and $\{\rho,0\}$, respectively. The subtree $\mathcal{T}_0$ of $\mathcal{T}$ consists of the nodes $\{0,00,01\}$, and its descendants $\mathcal{D}(\mathcal{T}_0)$ are $\{00,01\}$. The subtree $\mathcal{T}_\rho$ is the same as $\mathcal{T}$, and its descendants $\mathcal{D}(\mathcal{T}_\rho)$ are $\{0,1,00,01\}$. ◇

We now describe an algorithm that, for a given s.t. S, constructs a sequence of s.t.'s $\{T(\mathcal{T}^m)\}_{m=0}^M$ and Lagrange multipliers $\{\lambda_m\}_{m=0}^M$ with the following two properties. First, the $\mathcal{T}^m$'s and λ_m's are ordered in the sense that $\mathcal{T}^M\subset\mathcal{T}^{M-1}\subset\cdots\subset\mathcal{T}^0$ and $0=\lambda_M\leq\lambda_{M-1}\leq\cdots\leq\lambda_1\leq\lambda_0=\infty$. (Here, the symbol $\subset$ denotes inclusion, not necessarily strict.) Second, for any $m\in\{0,\ldots,M\}$ and $\lambda\in(\lambda_m,\lambda_{m-1}]$ the

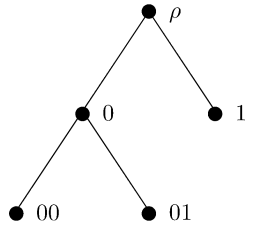

Fig. 3. Tree corresponding to the s.t. T defined by $T = 1$ if $Y_1 = 1$, and $T = 2$ else.

tree $\mathcal{T}^{m-1}$ minimizes $J_\lambda(\mathcal{T}) \triangleq J_\lambda(T(\mathcal{T}))$ among all nonrandomized s.t.'s. The algorithm builds upon ideas from the CART algorithm for the construction of classification and regression trees [4].

Before we state the algorithm, we need to introduce a few quantities. Given a nonrandomized s.t. T represented by its $|\mathcal{Y}|$-ary tree $\mathcal{T}$, we write the Lagrangian $J_\lambda(\mathcal{T})$ as

$$\begin{aligned} J_\lambda(\mathcal{T}) &= \mathbb{E}(T-S)^+ + \lambda \mathbb{P}(T<S) \\ &= \sum_{\boldsymbol{y}\in\mathcal{L}(\mathcal{T})} \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y})(\mathbb{E}((l(\boldsymbol{y})-S)^+|\boldsymbol{Y}=\boldsymbol{y}). \\ &\quad + \lambda \mathbb{P}(S>l(\boldsymbol{y})|\boldsymbol{Y}=\boldsymbol{y})) \\ &= \sum_{\boldsymbol{y}\in\mathcal{L}(\mathcal{T})} b(\boldsymbol{y}) + \lambda a(\boldsymbol{y}) \\ &= \sum_{\boldsymbol{y}\in\mathcal{L}(\mathcal{T})} J_\lambda(\boldsymbol{y}) \end{aligned}$$

where

$$\begin{aligned} a(\boldsymbol{y}) &\triangleq \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y})\mathbb{P}(S>l(\boldsymbol{y})|\boldsymbol{Y}=\boldsymbol{y}) \\ b(\boldsymbol{y}) &\triangleq \mathbb{P}(\boldsymbol{Y}=\boldsymbol{y})\mathbb{E}((l(\boldsymbol{y})-S)^+|\boldsymbol{Y}=\boldsymbol{y}) \\ J_\lambda(\boldsymbol{y}) &\triangleq b(\boldsymbol{y}) + \lambda a(\boldsymbol{y}). \end{aligned}$$

We extend the definition of $J_\lambda(\,\cdot\,)$ to subtrees of $\mathcal{T}$ by setting

$$J_\lambda(\mathcal{T}_{\boldsymbol{y}}) \triangleq \sum_{\boldsymbol{\gamma}\in\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})} J_\lambda(\boldsymbol{\gamma}).$$

With this definition[5]

$$J_\lambda(\mathcal{T}_{\boldsymbol{y}}) = \begin{cases} J_\lambda(\boldsymbol{y}), & \text{if } \boldsymbol{y}\in\mathcal{L}(\mathcal{T}) \\ \sum_{\gamma\in\mathcal{Y}} J_\lambda(\mathcal{T}_{\boldsymbol{y}\gamma}), & \text{if } \boldsymbol{y}\in\mathcal{I}(\mathcal{T}). \end{cases}$$

Similarly, we define

$$\begin{aligned} a(\mathcal{T}_{\boldsymbol{y}}) &\triangleq \sum_{\boldsymbol{\gamma}\in\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})} a(\boldsymbol{\gamma}), \\ b(\mathcal{T}_{\boldsymbol{y}}) &\triangleq \sum_{\boldsymbol{\gamma}\in\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})} b(\boldsymbol{\gamma}). \end{aligned}$$

For a given $\lambda \geq 0$ and $\mathcal{T}$, define $\mathcal{T}(\lambda) \subset \mathcal{T}$ to be the subtree of $\mathcal{T}$ such that $J_\lambda(\mathcal{T}(\lambda)) \leq J_\lambda(\mathcal{T}')$ for all subtrees , and $\mathcal{T}(\lambda) \subset \mathcal{T}'$ for all subtrees $\mathcal{T}' \subset \mathcal{T}$ satisfying $J_\lambda(\mathcal{T}(\lambda)) = J_\lambda(\mathcal{T}')$. In words, among all subtrees of $\mathcal{T}$ yielding a minimal cost for a given λ, the tree $\mathcal{T}(\lambda)$ is the smallest. As we shall see in Lemma 2, such a smallest subtree always exists, and hence $\mathcal{T}(\lambda)$ is well defined.

[5]We used $T, \mathcal{T}, \mathcal{T}_{\boldsymbol{y}}$, and $\boldsymbol{y}$ as possible arguments of $J_\lambda(\,\cdot\,)$. No confusion should arise from this slight abuse of notation, since for nonrandomized s.t.'s all of these arguments can be interpreted as trees.

Remark: $\mathcal{T}_{\boldsymbol{y}}(\lambda)$ is different from $(\mathcal{T}(\lambda))_{\boldsymbol{y}}$. Indeed, $\mathcal{T}_{\boldsymbol{y}}(\lambda)$ refers to the optimal subtree of $\mathcal{T}_{\boldsymbol{y}}$ with respect to λ, whereas $(\mathcal{T}(\lambda))_{\boldsymbol{y}}$ refers to subtree rooted in $\boldsymbol{y}$ of the optimal tree $\mathcal{T}(\lambda)$.

Example 6: Consider again the tree $\mathcal{T}$ in Fig. 3. Assume $J_\lambda(\rho) = 4, J_\lambda(0) = 2, J_\lambda(1) = J_\lambda(00) = J_\lambda(01) = 1$. Then

$$\begin{aligned} J_\lambda(\mathcal{T}) &= J_\lambda(1) + J_\lambda(00) + J_\lambda(01) = 3 \\ J_\lambda(\mathcal{T}_0) &= J_\lambda(00) + J_\lambda(01) = 2. \end{aligned}$$

The smallest optimal subtree of $\mathcal{T}$ is $\mathcal{T}(\lambda) = \{\rho, 0, 1\}$ and

$$J_\lambda(\mathcal{T}(\lambda)) = J_\lambda(0) + J_\lambda(1) = 3.$$

The smallest optimal subtree of $\mathcal{T}_0$ is $\mathcal{T}_0(\lambda) = \{0\}$ and

$$J_\lambda(\mathcal{T}_0(\lambda)) = J_\lambda(0) = 2. \qquad \diamond$$

Given a $|\mathcal{Y}|$-ary tree $\mathcal{T}$ and $\lambda \geq 0$, the following lemma shows that $\mathcal{T}(\lambda)$ always exists and characterizes $\mathcal{T}(\lambda)$ and $J_\lambda(\mathcal{T}(\lambda))$. The reader may recognize the Finite-Horizon backward induction algorithm whose detailed proof can be found in textbooks (e.g., [7, Chs. 3 and 4]).

Lemma 2: Pick a $|\mathcal{Y}|$-ary tree $\mathcal{T}$ and $\lambda \geq 0$. For every $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$

$$J_\lambda(\mathcal{T}_{\boldsymbol{y}}(\lambda)) = \min\left\{ J_\lambda(\boldsymbol{y}), \sum_{\gamma\in\mathcal{Y}} J_\lambda(\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda)) \right\}$$

and

$$\mathcal{T}_{\boldsymbol{y}}(\lambda) = \begin{cases} \{\boldsymbol{y}\}, & \text{if } J_\lambda(\boldsymbol{y}) \leq \sum_{\gamma\in\mathcal{Y}} J_\lambda(\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda)) \\ \{\boldsymbol{y}\} \cup_{\gamma\in\mathcal{Y}} \mathcal{T}_{\boldsymbol{y}\gamma}(\lambda), & \text{else.} \end{cases}$$

The optimal tree $\mathcal{T}(\lambda)$ and the corresponding cost $J_\lambda(\mathcal{T}(\lambda))$ are given by $J_\lambda(\mathcal{T}_{\boldsymbol{y}}(\lambda))$ and $\mathcal{T}_{\boldsymbol{y}}(\lambda)$ evaluated at $\boldsymbol{y} = \rho$.

Proof: By induction on the depth of the tree starting from the root. □

From the structure of the cost function $J_\lambda(\,\cdot\,)$, the larger the value of λ, the higher the penalty on the error probability. Therefore, one expects that the larger the λ the "later" the optimal tree $\mathcal{T}(\lambda)$ will stop. Indeed, Lemma 3 states that the tree corresponding to the optimal s.t. of a smaller λ is a subtree of the tree corresponding to the optimal s.t. of a larger λ. In other words, if $\lambda \leq \tilde{\lambda}$, in order to find $\mathcal{T}(\lambda)$, we can restrict our search to subtrees of $\mathcal{T}(\tilde{\lambda})$.

Lemma 3: Given a tree $\mathcal{T}$, if $\lambda \leq \tilde{\lambda}$ then $\mathcal{T}(\lambda) \subset \mathcal{T}(\tilde{\lambda})$.

Proof: We have

$$\begin{aligned} a(\mathcal{T}_{\boldsymbol{y}}) &= \sum_{\boldsymbol{y\gamma}\in\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})} \mathbb{P}(S>l(\boldsymbol{y\gamma})|Y^{l(\boldsymbol{y\gamma})}=\boldsymbol{y\gamma})\mathbb{P}(Y^{l(\boldsymbol{y\gamma})}=\boldsymbol{y\gamma}) \\ &\leq \sum_{\boldsymbol{y\gamma}\in\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})} \mathbb{P}(S>l(\boldsymbol{y})|Y^{l(\boldsymbol{y\gamma})}=\boldsymbol{y\gamma})\mathbb{P}(Y^{l(\boldsymbol{y\gamma})}=\boldsymbol{y\gamma}) \\ &= a(\boldsymbol{y}). \end{aligned} \tag{6}$$

Similarly, one shows that $b(\mathcal{T}_{\boldsymbol{y}}) \geq b(\boldsymbol{y})$.

By contradiction, assume $\lambda \leq \tilde{\lambda}$, but $\mathcal{T}(\lambda)$ is not a subtree of $\mathcal{T}(\tilde{\lambda})$. Then there exists $\boldsymbol{y} \in \mathcal{L}(\mathcal{T}(\tilde{\lambda}))$ such that $\boldsymbol{y} \in \mathcal{I}(\mathcal{T}(\lambda))$. By definition of $\mathcal{T}(\tilde{\lambda})$ and Lemma 2

$$J_{\tilde{\lambda}}(\boldsymbol{y}) \leq J_{\tilde{\lambda}}(\mathcal{T}_{\boldsymbol{y}}(\lambda))$$

and thus

$$b(\boldsymbol{y}) + \tilde{\lambda} a(\boldsymbol{y}) \leq b(\mathcal{T}_{\boldsymbol{y}}(\lambda)) + \tilde{\lambda} a(\mathcal{T}_{\boldsymbol{y}}(\lambda)). \tag{7}$$

Now, since $a(\mathcal{T}_{\boldsymbol{y}}(\lambda)) \leq a(\boldsymbol{y})$ and $\lambda \leq \tilde{\lambda}$

$$(\lambda - \tilde{\lambda}) a(\boldsymbol{y}) \leq (\lambda - \tilde{\lambda}) a(\mathcal{T}_{\boldsymbol{y}}(\lambda)). \tag{8}$$

Combining (7) and (8) yields

$$b(\boldsymbol{y}) + \lambda a(\boldsymbol{y}) \leq b(\mathcal{T}_{\boldsymbol{y}}(\lambda)) + \lambda a(\mathcal{T}_{\boldsymbol{y}}(\lambda))$$

and therefore

$$J_\lambda(\boldsymbol{y}) \leq J_\lambda(\mathcal{T}_{\boldsymbol{y}}(\lambda)).$$

Since $\boldsymbol{y} \in \mathcal{I}(\mathcal{T}(\lambda))$, this contradicts the definition of $\mathcal{T}(\lambda)$ by Lemma 2. □

The next theorem represents a key result. Given a tree $\mathcal{T}$, it characterizes the smallest value λ can take for which $\mathcal{T}(\lambda) = \mathcal{T}$. For a nontrivial tree $\mathcal{T}$, define for any $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$

$$g(\boldsymbol{y}, \mathcal{T}) \triangleq \frac{b(\mathcal{T}_{\boldsymbol{y}}) - b(\boldsymbol{y})}{a(\boldsymbol{y}) - a(\mathcal{T}_{\boldsymbol{y}})}$$

where we set $0/0 \triangleq 0$. The quantity $g(\boldsymbol{y}, \mathcal{T})$ captures the tradeoff between the reduction in delay $b(\mathcal{T}_{\boldsymbol{y}}) - b(\boldsymbol{y})$ and the increase in probability of false alarm $a(\boldsymbol{y}) - a(\mathcal{T}_{\boldsymbol{y}})$ if we stop at some intermediate node $\boldsymbol{y}$ instead of stopping at the leaves $\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})$ of $\mathcal{T}$.

Theorem 4: For any nontrivial tree $\mathcal{T}$

$$\inf\{\lambda \geq 0 : \mathcal{T}(\lambda) = \mathcal{T}\} = \max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T})} g(\boldsymbol{y}, \mathcal{T}).$$

Proof: Let $\mathcal{T}$ be a nontrivial tree and $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$. We have

$$\begin{aligned} g(\boldsymbol{y}, \mathcal{T}) &= \frac{J_\lambda(\mathcal{T}_{\boldsymbol{y}}) - \lambda a(\mathcal{T}_{\boldsymbol{y}}) - J_\lambda(\boldsymbol{y}) + \lambda a(\boldsymbol{y})}{a(\boldsymbol{y}) - a(\mathcal{T}_{\boldsymbol{y}})} \\ &= \frac{J_\lambda(\mathcal{T}_{\boldsymbol{y}}) - J_\lambda(\boldsymbol{y})}{a(\boldsymbol{y}) - a(\mathcal{T}_{\boldsymbol{y}})} + \lambda. \end{aligned}$$

By (6), $a(\mathcal{T}_{\boldsymbol{y}}) \leq a(\boldsymbol{y})$, and hence the following two implications hold:

$$\begin{aligned} g(\boldsymbol{y}, \mathcal{T}) \leq \lambda &\iff J_\lambda(\boldsymbol{y}) \geq J_\lambda(\mathcal{T}_{\boldsymbol{y}}) \\ g(\boldsymbol{y}, \mathcal{T}) < \lambda &\iff J_\lambda(\boldsymbol{y}) > J_\lambda(\mathcal{T}_{\boldsymbol{y}}). \end{aligned} \tag{9}$$

Therefore, if $\max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T})} g(\boldsymbol{y}, \mathcal{T}) < \lambda$ then

$$J_\lambda(\boldsymbol{y}) > J_\lambda(\mathcal{T}_{\boldsymbol{y}}) \tag{10}$$

for all $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$.

We first show by induction that if

$$\max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T})} g(\boldsymbol{y}, \mathcal{T}) < \lambda$$

then $\mathcal{T}(\lambda) = \mathcal{T}$. Consider a subtree of $\mathcal{T}$ having depth one and rooted in $\boldsymbol{y}$, say. Since $J_\lambda(\boldsymbol{y}) > J_\lambda(\mathcal{T}_{\boldsymbol{y}})$ by (10), we have $\mathcal{T}_{\boldsymbol{y}}(\lambda) = \mathcal{T}_{\boldsymbol{y}}$ by Lemma 2. Now consider a subtree of $\mathcal{T}$ with depth k, rooted in a different $\boldsymbol{y}$, and assume the assertion to be true for all subtrees of $\mathcal{T}$ with depth up to $k-1$. In order to find $\mathcal{T}_{\boldsymbol{y}}(\lambda)$, we use Lemma 2 and compare $J_\lambda(\boldsymbol{y})$ with $\sum_{\gamma \in \mathcal{Y}} J_\lambda(\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda))$. Since $\mathcal{T}_{\boldsymbol{y}\gamma}$ is a subtree of $\mathcal{T}$ with depth less than k, we have $\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda) = \mathcal{T}_{\boldsymbol{y}\gamma}$ by the induction hypothesis. Therefore

$$\sum_{\gamma \in \mathcal{Y}} J_\lambda(\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda)) = \sum_{\gamma \in \mathcal{Y}} J_\lambda(\mathcal{T}_{\boldsymbol{y}\gamma}) = J_\lambda(\mathcal{T}_{\boldsymbol{y}})$$

and since $J_\lambda(\mathcal{T}_{\boldsymbol{y}}) < J_\lambda(\boldsymbol{y})$ by (10), we have $\mathcal{T}_{\boldsymbol{y}}(\lambda) = \mathcal{T}_{\boldsymbol{y}}$ by Lemma 2, which concludes the induction step. Hence, we proved that if $\max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T})} g(\boldsymbol{y}, \mathcal{T}) < \lambda$, then $\mathcal{T}(\lambda) = \mathcal{T}$.

Second, suppose

$$\max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T})} g(\boldsymbol{y}, \mathcal{T}) = \lambda.$$

In this case, there exists $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$ such that $J_\lambda(\mathcal{T}_{\boldsymbol{y}}) = J_\lambda(\boldsymbol{y})$. We consider the cases when $\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda)$ and $\mathcal{T}_{\boldsymbol{y}\gamma}$ are the same for all $\gamma \in \mathcal{Y}$ and when they differ for at least one $\gamma \in \mathcal{Y}$. If $\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda) = \mathcal{T}_{\boldsymbol{y}\gamma}$ for all $\gamma \in \mathcal{Y}$ then

$$\sum_{\gamma \in \mathcal{Y}} J_\lambda(\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda)) = J_\lambda(\mathcal{T}_{\boldsymbol{y}}) = J_\lambda(\boldsymbol{y})$$

and thus $\mathcal{T}(\lambda) \neq \mathcal{T}$ by Lemma 2. If $\mathcal{T}_{\boldsymbol{y}\gamma}(\lambda) \neq \mathcal{T}_{\boldsymbol{y}\gamma}$ for at least one $\gamma \in \mathcal{Y}$ then $\mathcal{T}(\lambda) \neq \mathcal{T}$ again by Lemma 2.

Finally, if

$$\max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T})} g(\boldsymbol{y}, \mathcal{T}) > \lambda$$

then $\mathcal{T}(\lambda) \neq \mathcal{T}$ follows from the previous case and Lemma 3. □

Let $\mathcal{T}^0$ denote the complete tree of depth κ. Starting with $\lambda_0 = \infty$, for $m = \{1, \ldots, M\}$ recursively define

$$\begin{aligned} \lambda_m &\triangleq \inf\{\lambda \leq \lambda_{m-1} : \mathcal{T}^{m-1}(\lambda) = \mathcal{T}^{m-1}\} \\ \mathcal{T}^m &\triangleq \mathcal{T}^{m-1}(\lambda_m) \end{aligned}$$

where M is the smallest integer such that $\lambda_{M+1} = 0$, and with $\lambda_1 \triangleq \infty$ if the set over which the infimum is taken is empty. Lemma 3 implies that for two consecutive transition points λ_m and λ_{m+1}, we have $\mathcal{T}^0(\lambda) = \mathcal{T}^0(\lambda_m)$ for all $\lambda \in (\lambda_{m+1}, \lambda_m]$ as shown in Fig. 4.

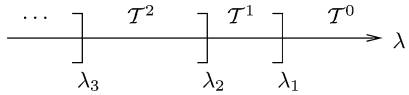

Fig. 4. For all $m \in \{0, 1, \ldots, M-1\}$ the tree $\mathcal{T}^m$ is the smallest tree minimizing the cost $J_\lambda(\cdot)$ for any $\lambda \in (\lambda_{m+1}, \lambda_m]$.

The following corollary is a consequence of Lemma 3 and Theorem 4.

Corollary 5: For $m \in \{1, \ldots, M\}$

$$\lambda_m = \max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^{m-1})} g(\boldsymbol{y}, \mathcal{T}^{m-1}) \tag{11}$$

$$\mathcal{T}^m = \mathcal{T}^{m-1} \setminus \bigcup_{\substack{\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^{m-1}): \\ g(\boldsymbol{y}, \mathcal{T}^{m-1}) = \lambda_m}} \mathcal{D}\left(\mathcal{T}^{m-1}_{\boldsymbol{y}}\right). \tag{12}$$

Moreover, the set $\{(\alpha_m, d_m)\}_{m=1}^M$ with

$$\begin{aligned} \alpha_m &\triangleq \mathbb{P}(T(\mathcal{T}^m) < S), \\ d_m &\triangleq \mathbb{E}(T(\mathcal{T}^m) - S)^+ \end{aligned}$$

are the breakpoints of $d(\alpha)$.

Proof: Let $\mathcal{T}^{m-1}$ be fixed. Equation (11) follows directly from Theorem 4. For (12), notice that as $J_\lambda(\mathcal{T})$ is continuous

in λ, the definition of λ_m yields $J_{\lambda_m}(\mathcal{T}^{m-1}) = J_{\lambda_m}(\mathcal{T}^m)$. Hence, $\mathcal{T}^m$ is the smallest subtree of $\mathcal{T}^{m-1}$ having a cost equal to $J_{\lambda_m}(\mathcal{T}^{m-1})$. From (9) and Lemma 2, we deduce that $\mathcal{T}^m$ is obtained from $\mathcal{T}^{m-1}$ by removing the descendants of any $\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^{m-1})$ such that $g(\boldsymbol{y}, \mathcal{T}^{m-1}) = \lambda_m$.

It remains to show that $\{(\alpha_m, d_m)\}_{m=1}^M$ are the breakpoints of $d(\alpha)$. By Lemma 1, the breakpoints are achieved by nonrandomized s.t.'s. By Lemma 3, we have $\mathcal{T}^m = \mathcal{T}^0(\lambda_m)$, i.e., $\mathcal{T}^m$ is the smallest subtree of $\mathcal{T}^0$ minimizing the cost $J_{\lambda_m}(\mathcal{T})$. Hence, among the minimizers of $J_{\lambda_m}(\mathcal{T})$, $\mathcal{T}^m$ yields the largest $\mathbb{P}(T(\mathcal{T}) < S)$. Therefore, each pair (α_m, d_m) is a breakpoint. Conversely, given a breakpoint of $d(\alpha)$, let $\mathcal{T}$ be the smallest subtree of $\mathcal{T}^0$ achieving it. Then $\mathcal{T} = \mathcal{T}^0(\lambda)$ for some λ. Since $\mathcal{T}^0(\lambda_m) = \mathcal{T}^m$ we have that $\{\mathcal{T}^0(\lambda)\}_{\lambda \in \mathbb{R}} = \{\mathcal{T}^m\}_{m=0}^M$, and, therefore, $\mathcal{T} = \mathcal{T}^m$ for some $m \in \{1, \ldots, M\}$. □

From Corollary 5, we deduce the algorithm below that fully characterizes $d(\alpha)$ by computing its set of breakpoints $\{(\alpha_m, d_m)\}_{m=1}^M$.

Algorithm: Compute breakpoints $\{(\alpha_m, d_m)\}_{m=1}^M$ of $d(\alpha)$

$m \Leftarrow 0$

$\lambda_0 \Leftarrow \infty$

$\mathcal{T}^0 \Leftarrow$ complete tree of depth κ

Repeat

$m \Leftarrow m + 1$

$\lambda_m \Leftarrow \max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^{m-1})} g(\boldsymbol{y}, \mathcal{T}^{m-1})$

$\mathcal{T}^m \Leftarrow \mathcal{T}^{m-1} \setminus \bigcup_{\substack{\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^{m-1}): \\ g(\boldsymbol{y}, \mathcal{T}^{m-1}) = \lambda_m}} \mathcal{D}(\mathcal{T}_{\boldsymbol{y}}^{m-1})$

$\alpha_m \Leftarrow \mathbb{P}(T(\mathcal{T}^m) < S)$

$d_m \Leftarrow \mathbb{E}(T(\mathcal{T}^m) - S)^+$

until $\lambda_m = 0$

$M \Leftarrow m - 1$

As a $|\mathcal{Y}|$-ary tree has less than $|\mathcal{Y}|^\kappa$ nonterminal nodes, the algorithm terminates after at most that many iterations. Further, one may check that each iteration has a running time that is $\exp(O(\kappa))$. Therefore, the worst case running time of the algorithm is $\exp(O(\kappa))$. This is to be compared, for instance, with exhaustive search, which has a $\Omega(\exp(\exp(\kappa)))$ running time (because all breakpoints of $d(\alpha)$ are achieved by nonrandomized s.t.'s and there are already $2^{|\mathcal{Y}|^{\kappa-1}}|\mathcal{Y}|$-ary trees having leaves at either depth κ or $\kappa - 1$).

In Sections IV and V we will see that, under certain conditions on $\{(X_i, Y_i)\}_{i \geq 1}$ and S, the running time of the algorithm is only *polynomial* in κ.

A. A Lower Bound on the Reaction Delay

From Corollary 5, we may also deduce a lower bound on $d(\alpha)$. Since $d(\alpha)$ is convex, we can lower-bound it as

$$d(\alpha) \geq d(0) + \alpha d'(0+) \tag{13}$$

where $d'(0+)$ denotes the right derivative of d at $\alpha = 0$. By Corollary 5, if $\lambda_1 < \infty$ then $d(0)$ is achieved by the complete tree $\mathcal{T}^0$, and if $\lambda_1 = \infty$ then $d(0)$ is achieved by $\mathcal{T}^1$ which is a strict subtree of $\mathcal{T}^0$. Hence, (13) can be written as

$$d(\alpha) \geq d(0) - \begin{cases} \alpha\lambda_1, & \text{if } \lambda_1 < \infty \\ \alpha\lambda_2, & \text{else.} \end{cases} \tag{14}$$

the above bound is tight for $\alpha \leq \alpha_1$ with $\alpha_1 > 0$ when $\lambda_1 < \infty$, and is tight for $\alpha \leq \alpha_2$ with $\alpha_2 > 0$ when $\lambda_1 = \infty$. The following example illustrates this bound.

Example 7: Let $\{X_i\}_{i \geq 1}$ be independent and identically distributed (i.i.d.) Bernoulli $(1/2)$, and let the Y_i's be the output of a binary-symmetric channel with crossover probability $p \in (0, 1/2)$ for input X_i. Consider the s.t. S defined as

$$S \triangleq \begin{cases} 1, & \text{if } X_1 = 1 \\ \kappa, & \text{else.} \end{cases}$$

For $\kappa = 2$, the tree corresponding to this s.t. is depicted in Fig. 3.

Since $p \in (0, 1/2)$, it is clear that whenever $\mathcal{T}$ is not the complete tree of depth κ, we have $\mathbb{P}(T(\mathcal{T}) < S) > 0$, hence

$$d(0) = \mathbb{E}(T(\mathcal{T}^0) - S)^+ = \frac{1}{2}(\kappa - 1).$$

An easy computation using Corollary 5 yields

$$\lambda_1 = \frac{1-p}{p}(\kappa - 1),$$

and using (14), we get

$$d(\alpha) \geq (\kappa - 1)\left(\frac{1}{2} - \alpha\frac{1-p}{p}\right). \tag{15}$$

Let us comment on (15). Consider any two correlated sequences $\{X_i\}_{i \geq 1}$ and $\{Y_i\}_{i \geq 1}$ and a s.t. S with respect to the X_i's. Intuition tells us that there are two factors affecting $d(\alpha)$. The first is the correlation between the X_i's and Y_i's, in the above example parameterized by p. The lower the correlation, the higher $d(\alpha)$. The second factor is the "variability" of S, and might be characterized by the difference in terms of depth among the leaves having large probability to be reached. In the above example the "variability" might be captured by $\kappa - 1$, since with probability $1/2$ a leaf of depth 1 is reached, and with probability $1/2$ a leaf of depth κ is attained. □

Example 8.: We refer to the feedback communication problem of Example 2. Consider the "send until non-erasure" strategy when the forward and the feedback channels are binary erasure channels with erasure probabilities ε and p, respectively. The decoder keeps sending 0's over the feedback channel until time S, the first time a non-erasure occurs or κ time units have elapsed (the parameter κ plays the role of a "time-out"). From that time on the decoder sends 1's. The encoder keeps sending the message bit it wants to deliver until time T (a s.t. with respect to the output of the feedback channel).We analyze the communication rate with respect to $\mathbb{E}(\max\{S, T\})$. To that aim, we investsigate $d(\alpha)$.

One can show that $\lambda_1 = \infty$ and, therefore, the bound (14) becomes $d(\alpha) \geq d(0) - \alpha\lambda_2$, where $\lambda_2 = \max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^1)} g(\boldsymbol{y}, \mathcal{T}^1)$ from Corollary 5. A somewhat involved computation yields

$$d(\alpha) \geq \left(\frac{p}{1-p} - \varepsilon^{1-\kappa}\alpha\right)(1 + o(1)) \tag{16}$$

as $\kappa \to \infty$.

For the communication scheme considered here, there are two events leading to decoding errors. The event $\{X_\kappa = 0\}$, indicating that only erasures were received by the decoder until time κ, and the event $\{T < S\}$, indicating that the encoder stopped retransmission before the decoder received a non-erasure. In both cases, the decoder will make an error with probability $1/2$. Hence, the probability of error $\mathbb{P}(\mathcal{E})$ can be bounded as

$$\max\{\alpha, \varepsilon^\kappa\} \leq 2\mathbb{P}(\mathcal{E}) \leq \alpha + \varepsilon^\kappa.$$

It is then reasonable to choose $\kappa = \frac{\log \alpha}{\log \varepsilon}$, i.e., to scale κ with α so that both sources of errors have the same weight. This results in a minimum reaction delay of

$$d(\alpha) \geq \left(\frac{p}{1-p} - \varepsilon\right)(1 + o(1))$$

as $\alpha \to 0$.

Since the communication rate R is computed with respect to $\mathbb{E}(\max\{S,T\}) = \mathbb{E}S + \mathbb{E}(T-S)^+$. We conclude that the "send until a non-erasure" strategy asymptotically achieves a rate that is upper-bounded as

$$R \leq \frac{1}{\frac{1}{1-\varepsilon} + \frac{p}{1-p} - \varepsilon}.$$

When $\varepsilon < p/(1-p)$, this bound is strictly below the capacity of the binary erasure channel $1-\varepsilon$. Hence, $1/(1+\varepsilon)$ represents a critical value for the erasure probability p of the feedback channel above which the "send until non-erasure" strategy is strictly suboptimal. Indeed, there exist block coding strategies, making no use of feedback, that (asymptotically) achieve rates up to $1-\varepsilon$, the capacity of the forward channel. $\diamond$

IV. Permutation-Invariant Stopping Times

We consider a special class of s.t.'s and processes $\{(X_i, Y_i)\}_{i\geq 1}$ for which the optimal tradeoff curve $d(\alpha)$ and the associated optimal s.t.'s can be computed in polynomial time in κ.

A s.t. S with respect to $\{X_i\}_{i\geq 1}$ is *permutation invariant* if

$$\mathbb{P}(S \leq n | X^n = x^n) = \mathbb{P}(S \leq n | X^n = \pi(x^n))$$

for all permutations $\pi : \mathcal{X}^n \to \mathcal{X}^n$, all $x^n \in \mathcal{X}^n$, and $n \in \{1, \ldots, \kappa\}$. Examples of permutation-invariant s.t.'s are $S = \inf\{i : X_i > c\}$ or $S = \inf\{i : \sum_{k=1}^i X_k > c\}$ for some constant c and assuming the X_i's are positive. The notion of a permutation-invariant s.t. is closely related to (and in fact slightly stronger than) that of an exchangeable s.t. as defined in [15].

The following theorem establishes a key result, from which the running time of one iteration of the algorithm can be deduced.

Theorem 6: Let $\{(X_i, Y_i)\}_{i\geq 1}$ be i.i.d. and S be a permutation-invariant s.t. with respect to $\{X_i\}_{i\geq 1}$. If $T(\mathcal{T})$ is nonrandomized and permutation invariant then

$$g(\boldsymbol{y}, \mathcal{T}) = g(\pi(\boldsymbol{y}), \mathcal{T})$$

for all $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$ and all permutations π.

We first establish two lemmas that will be used in the proof of Theorem 6.

Lemma 7: Let T be a nonrandomized s.t. with respect to $\{Y_i\}_{i\geq 1}$ and $\mathcal{T}$ the corresponding tree. Then T is permutation invariant if and only if for all $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$ and permutations $\pi, \pi(\boldsymbol{y}) \in \mathcal{I}(\mathcal{T})$.

Proof: Assume T is permutation invariant and let $y^n \in \mathcal{I}(\mathcal{T})$. Then

$$0 = \mathbb{P}(T \leq n | Y^n = y^n) = \mathbb{P}(T \leq n | Y^n = \pi(y^n)),$$

and hence $\pi(y^n) \in \mathcal{I}(\mathcal{T})$.

Conversely assume that, for all $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$ and permutations π, we have $\pi(\boldsymbol{y}) \in \mathcal{I}(\mathcal{T})$. Pick an arbitrary y^n. First, if $\mathbb{P}(T \leq n | Y^n = y^n) = 0$, then $y^n \in \mathcal{I}(\mathcal{T})$, and by assumption also $\pi(y^n) \in \mathcal{I}(\mathcal{T})$. Thus, $\mathbb{P}(T \leq n | Y^n = \pi(y^n)) = 0$. Second, if $\mathbb{P}(T \leq n | Y^n = y^n) = 1$, then $y^n \notin \mathcal{I}(\mathcal{T})$, and by assumption also $\pi(y^n) \notin \mathcal{I}(\mathcal{T})$. Thus, $\mathbb{P}(T \leq n | Y^n = \pi(y^n)) = 1$. $\square$

Lemma 8: Let $\{(X_i, Y_i)\}_{i\geq 1}$ be i.i.d. and S be a permutation-invariant s.t. with respect to $\{X_i\}_{i\geq 1}$. Then S is a permutation-invariant s.t. with respect to $\{Y_i\}_{i\geq 1}$.

Proof: Using that the $\{(X_i, Y_i)\}_{i\geq 1}$ are i.i.d., one can easily check that S is a s.t. with respect to $\{Y_i\}_{i\geq 1}$. It remains to show that it is permutation invariant. For any permutation $\pi : \mathcal{X}^n \to \mathcal{X}^n$

$$\begin{aligned}
&\mathbb{P}(S \leq n | Y^n = y^n) \\
&\quad = \sum_{x^n \in \mathcal{X}^n} \mathbb{P}(S \leq n | X^n = x^n) \mathbb{P}(X^n = x^n | Y^n = y^n) \\
&\quad = \sum_{x^n \in \mathcal{X}^n} \mathbb{P}(S \leq n | X^n = \pi^{-1}(x^n)) \times \\
&\qquad \times \mathbb{P}(X^n = \pi^{-1}(x^n) | Y^n = y^n) \\
&\quad = \sum_{x^n \in \mathcal{X}^n} \mathbb{P}(S \leq n | X^n = x^n) \mathbb{P}(X^n = x^n | Y^n = \pi(y^n)) \\
&\quad = \mathbb{P}(S \leq n | Y^n = \pi(y^n))
\end{aligned}$$

where the second last equality follows by the permutation invariance of S and the fact that the (X_i, Y_i)'s are i.i.d. $\square$

Proof of Theorem 6: We show that

$$g(\boldsymbol{y}, \mathcal{T}) = \frac{b(\mathcal{T}_{\boldsymbol{y}}) - b(\boldsymbol{y})}{a(\boldsymbol{y}) - a(\mathcal{T}_{\boldsymbol{y}})} = g(\pi(\boldsymbol{y}), \mathcal{T}) \tag{17}$$

for all $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$. We prove that the numerator and the denominator in (17) remain unchanged if we replace $\boldsymbol{y}$ by $\pi(\boldsymbol{y})$. Fix some $\boldsymbol{y} = y^n \in \mathcal{I}(\mathcal{T})$, and, to simplify notation, set $l = l(\boldsymbol{\gamma})$ until the end of this proof. For the denominator, using Lemma 8 we obtain

$$\begin{aligned}
&a(\boldsymbol{y}) - a(\mathcal{T}_{\boldsymbol{y}}) \\
&\quad \triangleq a(y^n) - \sum_{y^n\boldsymbol{\gamma} \in \mathcal{L}(\mathcal{T}_{y^n})} a(y^n\boldsymbol{\gamma}) \\
&\quad = \mathbb{P}(Y^n = y^n)\mathbb{P}(S > n | Y^n = y^n) \\
&\qquad - \sum_{y^n\boldsymbol{\gamma} \in \mathcal{L}(\mathcal{T}_{y^n})} \mathbb{P}(Y^{n+l} = y^n\boldsymbol{\gamma}) \mathbb{P}(S > n+l | Y^{n+l} = y^n\boldsymbol{\gamma}) \\
&\quad = \mathbb{P}(Y^n = \pi(y^n))\mathbb{P}(S > n | Y^n = \pi(y^n))
\end{aligned}$$

$$- \sum_{y^n\boldsymbol{\gamma}\in\mathcal{L}(\mathcal{T}_{y^n})} \mathbb{P}(Y^{n+l} = \pi(y^n)\boldsymbol{\gamma})\times \\ \times\ \mathbb{P}(S > n+l | Y^{n+l} = \pi(y^n)\boldsymbol{\gamma}). \tag{18}$$

A consequence of Lemma 7 is that the set of all $\boldsymbol{\gamma}$ such that $y^n\boldsymbol{\gamma} \in \mathcal{L}(\mathcal{T}_{y^n})$ is identical to the set of all $\boldsymbol{\gamma}$ such that $\pi(y^n)\boldsymbol{\gamma} \in \mathcal{L}(\mathcal{T}_{\pi(y^n)})$. Hence, by (18)

$$a(y^n) - a(\mathcal{T}_{y^n}) = a(\pi(y^n)) - a(\mathcal{T}_{\pi(y^n)}).$$

For the numerator in (17), we have

$$b(\mathcal{T}_{y^n}) - b(y^n) \\ = \sum_{y^n\boldsymbol{\gamma}\in\mathcal{L}(\mathcal{T}_{y^n})} \mathbb{P}(Y^{n+l} = y^n\boldsymbol{\gamma})\times \\ \times \Big(\mathbb{E}((n+l-S)^+ | Y^{n+l} = y^n\boldsymbol{\gamma}) \\ - \mathbb{E}((n-S)^+ | Y^{n+l} = y^n\boldsymbol{\gamma})\Big). \tag{19}$$

By Lemma 8

$$\mathbb{E}((n+l-S)^+ | Y^{n+l} = y^n\boldsymbol{\gamma}) - \mathbb{E}((n-S)^+ | Y^{n+l} = y^n\boldsymbol{\gamma}) \\ = \sum_{k=n}^{n+l-1} \mathbb{P}(S \le k | Y^{n+l} = y^n\boldsymbol{\gamma}) \\ = \sum_{k=n}^{n+l-1} \mathbb{P}(S \le k | Y^{n+l} = \pi(y^n)\boldsymbol{\gamma}).$$

Combining this with (19) and using Lemma 7 as before, we get

$$b(\mathcal{T}_{y^n}) - b(y^n) = b(\mathcal{T}_{\pi(y^n)}) - b(\pi(y^n))$$

concluding the proof. □

We now show that one iteration of the algorithm has only polynomial running time in κ. Specifically, we evaluate the running time to compute $\mathcal{T}^{m+1}$ from $\mathcal{T}^m$ if S and $T(\mathcal{T}^m)$ are permutation invariant and if the (X_i, Y_i)'s are i.i.d. To that aim, we assume the input of the algorithm to be in the form of a list of the probabilities $\mathbb{P}(S \le n | X^n = x^n)$ for all $x^n \in \mathcal{X}^n$ and $n \in \{1, \dots, \kappa\}$—specifying S—and a list of $\mathbb{P}(X = x, Y = y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$—characterizing the distribution of the process $\{(X_i, Y_i)\}_{i\ge 1}$. As S is permutation invariant, we only have to specify $\mathbb{P}(S \le n | X^n = x^n)$ for each composition[6] of x^n. Since the number of compositions of length at most κ is upper-bounded by $(\kappa+1)^{1+|\mathcal{X}|}$—any element $x \in \mathcal{X}$ appears at most k times in a string of length k—the list of these probabilities has only polynomial size in κ. Using a hash table, we assume that, given x^n, the element $\mathbb{P}(S \le n | X^n = x^n)$ in the list can be accessed in $O(\kappa)$ time. The proof of the following theorem is deferred to Appendix I.

Theorem 9: Let $\{(X_i, Y_i)\}_{i\ge 1}$ be i.i.d., let S and $T(\mathcal{T}^m)$ be permutation-invariant s.t.'s with respect to $\{X_i\}_{i\ge 1}$ and $\{Y_i\}_{i\ge 1}$, respectively, and let $\alpha_m = \mathbb{P}(T(\mathcal{T}^m) < S)$ and $d_m = \mathbb{E}(T(\mathcal{T}^m) - S)^+$ be given. Then, $\mathcal{T}^{m+1}, \alpha_{m+1}$, and d_{m+1} can be computed in polynomial time in κ.

[6]Two sequences have the same composition (or type) if any particular symbol appears the same number of times in both the sequences.

As a corollary of Theorem 9, we obtain the worst case running time for computing the set of breakpoints $\{(\alpha_m, d_m)\}_{m=1}^M$ together with the associated optimal s.t.'s $\{\mathcal{T}^m\}_{m=0}^M$.

Corollary 10: Let $\{(X_i, Y_i)\}_{i\ge 1}$ be i.i.d. and S be a permutation-invariant s.t. with respect to $\{X_i\}_{i\ge 1}$. If all $\{\mathcal{T}^m\}_{m=0}^M$ are permutation invariant, then the algorithm has a polynomial running time in κ.

Proof: By Theorem 9 we only have to bound the number of iterations of the algorithm. To this end note that by Theorem 6 every composition of $\boldsymbol{y}$ can be only once a maximizer of $g(\boldsymbol{y}, \mathcal{T}^m)$ (as the corresponding nodes will be leaves in the next iteration of the algorithm). Hence, there are at most $O((\kappa+1)^{1+|\mathcal{Y}|})$ iterations. □

Note that, in the cases where $\{\mathcal{T}^m\}_{m=0}^M$ are not permutation invariant, one may still be able to derive a lower bound on $d(\alpha)$ in polynomial time in κ, using (14). Indeed, the tree $\mathcal{T}^0$ is permutation invariant since it is complete and, by Theorem 9, if $\{(X_i, Y_i)\}_{i\ge 1}$ are i.i.d. and S is permutation invariant, then the first subtree $\mathcal{T}^1$ can be computed in polynomial time in κ. Therefore, the bound

$$d(\alpha) \ge d(0) - \alpha\lambda_1 \tag{20}$$

can always be evaluated in polynomial time in κ when the (X_i, Y_i)'s are i.i.d. and S is permutation invariant. this bound is in general weaker than the one derived in Section III-A. However, when $\lambda_1 < \infty$ the bound (20) is tight for $\alpha \in [0, \alpha_1]$ with $\alpha_1 > 0$. It is easily checked that the condition $\lambda_1 < \infty$ is satisfied if $\mathbb{P}(S = \kappa, Y^{\kappa-1} = y^{\kappa-1}) > 0$ for all $y^{\kappa-1}$.

In the next section, we present two examples for which the conditions of Corollary 10 are satisfied, and hence for which the algorithm has a polynomial running time in κ. First, we consider a TST problem that indeed can be formulated as a Bayesian change-point problem. Second, we consider the case of a pure TST problem, i.e., one that cannot be formulated as a Bayesian change-point problem. For both examples, we provide an analytical solution of the Lagrange minimization problem $\min_{T\le\kappa} J_\lambda(T)$.

V. One-Step Look-Ahead Stopping Times

Define

$$\mathcal{A}_n \triangleq \left\{ y^n \in \mathcal{Y}^n : \sum_{\gamma\in\mathcal{Y}} J_\lambda(y^n\gamma) \ge J_\lambda(y^n) \right\}$$

and let

$$T_\lambda^* \triangleq \min\left\{\kappa, \inf\{n : Y^n \in \mathcal{A}_n\}\right\}. \tag{21}$$

In words, T_λ^* stops whenever the current cost

$$\mathbb{E}((n-S)^+ | Y^n = y^n) + \lambda\mathbb{P}(S > n | Y^n = y^n)$$

is less than the expected cost at time $n+1$, i.e.,

$$\mathbb{E}((n+1-S)^+ | Y^n = y^n) + \lambda\mathbb{P}(S > n+1 | Y^n = y^n).$$

Recall that $\mathcal{T}^0$ denotes the complete tree of depth κ. For (X_i, Y_i)'s i.i.d., Theorem 11 provides a sufficient condition on S for which $T(\mathcal{T}^0(\lambda)) = T_\lambda^*$. In words, the s.t. T_λ^* minimizes $J_\lambda(T)$ among all s.t.'s bounded by κ. Furthermore, among all s.t.'s minimizing $J_\lambda(T)$, the s.t. T_λ^* admits the smallest

tree representation. The proof of Theorem 11 is deferred to Appendix II.

Theorem 11: Let $\{(X_i, Y_i)\}_{i\geq 1}$ be i.i.d., and let S be a s.t. with respect to $\{X_i\}_{i\geq 1}$ that satisfies

$$\mathbb{P}(S = n|Y^{n-1}) \geq \mathbb{P}(S = n+1|Y^n) \tag{22}$$

for all $n \in \{2, \ldots, \kappa\}$. Then

$$T(\mathcal{T}^0(\lambda)) = T^*_\lambda.$$

Note that, unlike the algorithm, Theorem 11 provides an analytical solution only to the inner minimization problem in (5). To find the reaction delay $d(\alpha)$ one still needs to maximize over the Lagrange multipliers λ.

Using Theorems 10 and 11, we now give two examples of process $\{(X_i, Y_i)\}_{i\geq 1}$ and s.t. S for which the algorithm has only polynomial running time in κ.

Example 9: Let $\{(X_i, Y_i)\}_{i\geq 1}$ be i.i.d. with the X_i's taking values in $\{0, 1\}$. Consider the s.t. $S \triangleq \inf\{i : X_i = 1\}$. We have for $n \geq 2$

$$\begin{aligned}
&\mathbb{P}(S = n|Y^{n-1}) \\
&\quad = \mathbb{P}(S \geq n|Y^{n-1})\mathbb{P}(X_n = 1) \\
&\quad \geq \mathbb{P}(S \geq n|Y^{n-1})\mathbb{P}(X_n = 0|Y_n)\mathbb{P}(X_{n+1} = 1) \\
&\quad = \mathbb{P}(S = n+1|Y^n).
\end{aligned}$$

Hence, Theorem 11 yields that the one-step look-ahead s.t. T^*_λ defined in (21) satisfies $T(\mathcal{T}^0(\lambda)) = T^*_\lambda$.

We now show that the algorithm finds the set of breakpoints $\{(\alpha_m, d_m)\}_{m=0}^M$ and the corresponding $\{\mathcal{T}_m\}_{m=0}^M$ in polynomial running time in κ. To that aim, we first show that T^*_λ is permutation invariant. By Lemma 7, we equivalently show that, for all y^n and permutations π, if $y^n \notin \mathcal{A}_n$ then $\pi(y^n) \notin \mathcal{A}_n$. We have for $n < \kappa$

$$\begin{aligned}
&\sum_{\gamma\in\mathcal{Y}} J_\lambda(y^n\gamma) - J_\lambda(y^n) \\
&\quad = \mathbb{P}(Y^n = y^n)\Big(\mathbb{P}(S \leq n|Y^n = y^n) \\
&\qquad - \lambda\mathbb{P}(S = n+1|Y^n = y^n)\Big) \\
&\quad = \mathbb{P}\Big(Y^n = \pi(y^n))(\mathbb{P}(S \leq n|Y^n = \pi(y^n)) \\
&\qquad - \lambda\mathbb{P}(S = n+1|Y^n = \pi(y^n))\Big) \\
&\quad = \sum_{\gamma\in\mathcal{Y}} J_\lambda(\pi(y^n)\gamma) - J_\lambda(\pi(y^n))
\end{aligned} \tag{23}$$

where we have used Lemma 8 for the second equality. Thus, $y^n \notin \mathcal{A}_n$ implies $\pi(y^n) \notin \mathcal{A}_n$, and therefore T^*_λ is permutation invariant. Since $T(\mathcal{T}^0(\lambda)) = T^*_\lambda$ for all $\lambda \geq 0$ by Theorem 11, all $\{\mathcal{T}^m\}_{m=0}^M$ are permutation invariant. Finally, because S is permutation invariant, applying Corollary 10 we conclude that the algorithm has indeed polynomial running time in κ.

The problem considered in this example is actually a Bayesian change-point problem, as defined in Example 4 in Section I. Here, the change-point $\Theta \triangleq S$ has distribution $\mathbb{P}(\Theta = n) = p(1-p)^{n-1}$, where $p \triangleq \mathbb{P}(X = 1)$. The conditional distribution of Y_i given Θ is

$$\mathbb{P}(Y_i = y_i|\Theta = n) = \begin{cases} \mathbb{P}(Y_i = y_i|X_i = 0), & \text{if } i < n \\ \mathbb{P}(Y_i = y_i|X_i = 1), & \text{if } i = n \\ \mathbb{P}(Y_i = y_i), & \text{if } i > n. \end{cases}$$

unlike the case considered by Shiryaev (see Example 4 in Section I), the distribution of the process at the change-point differs from the ones before and after it. ◇

We now give an example that cannot be formulated as a change-point problem and for which the one-step look-ahead s.t. T^*_λ minimizes the Lagrangian $J_\lambda(T)$.

Example 10: Let $\{(X_i, Y_i)\}_{i\geq 1}$ be i.i.d. where the X_i's and Y_i's take values in $\{0, 1\}$, and let $S \triangleq \inf\{i \geq 1 : \sum_{j=1}^i X_j = 2\}$. A similar computation as for Example 9 reveals that if

$$\mathbb{P}(X_i = 1|Y_i) \geq \mathbb{P}(X_i = 0|Y_i)$$

then Theorem 11 applies, showing that the one-step look-ahead s.t. T^*_λ defined in (21) satisfies $T(\mathcal{T}^0(\lambda)) = T^*_\lambda$.

Furthermore, since S is permutation invariant, (23) shows that T^*_λ is permutation invariant. Applying Corollary 10, one deduces that the algorithm has polynomial running time in κ in this case as well.

The problem considered here is *not* a change-point problem since, for $k > n$

$$\mathbb{P}(S = k|Y^n = y^n, S > n) \neq \mathbb{P}(S = k|S > n)$$

and therefore (1) does not hold. ◇

VI. Remarks

In our analysis, we exploited the finite tree structure of bounded s.t.'s defined over finite alphabet processes, and derived an algorithm that outputs the minimum reaction delays for tracking a s.t. through noisy observations, for any probability of false alarm. This algorithm has a complexity that is exponential in the bound of the s.t. we want to track and, in certain cases, even polynomial. In comparison, an exhaustive search has a complexity that is doubly exponential.

The conditions under which the algorithm runs in polynomial time are, unfortunately, not very explicit and require further study (see Corollary 10). Explicit conditions, however, are expected to be very restrictive on both the stochastic process and the s.t. to be tracked.

For certain applications, it is suitable to consider s.t.'s defined over more general processes, such as continuous time over continuous alphabets. In this case, how to solve the TST problem remains a wide open question. As a first step, one might consider a time and alphabet quantization and apply our result in order to derive an approximation algorithm.

Appendix I
Proof of Theorem 9

In the following, we write $\mathcal{T}$ for $\mathcal{T}^m$. From Theorem 6, to find the $\boldsymbol{y} \in \mathcal{I}(\mathcal{T})$ maximizing $g(\boldsymbol{y}, \mathcal{T})$, we only have to compute $g(\boldsymbol{y}, \mathcal{T})$ for all possible compositions of $\boldsymbol{y}$. The number of such

compositions is $O((\kappa+1)^{1+|\mathcal{Y}|})$. We now show that $g(\boldsymbol{y},\mathcal{T})$ can be computed in polynomial time in κ. From the Proof of Theorem 6, we have to show that $\mathbb{P}(S \le n|Y^n = y^n)$ can be computed in polynomial time, and that the sums in (18) and (19) can be computed in polynomial time.

We have

$$\mathbb{P}(S \le n|Y^n = y^n) = \sum_{x^n \in \mathcal{X}^n} \mathbb{P}(S \le n|X^n = x^n)\mathbb{P}(X^n = x^n|Y^n = y^n).$$

Each term in the summation on the right-hand side depends only on the composition of (x^n, y^n), and hence $\mathbb{P}(S \le n|Y^n = y^n)$ can be computed in polynomial time in κ.

Consider now the sum over all $y^n\boldsymbol{\gamma} \in \mathcal{L}(\mathcal{T}_{y^n})$ in (18)

$$\sum_{y^n\boldsymbol{\gamma}\in\mathcal{L}(\mathcal{T}_{y^n})} a(y^n\boldsymbol{\gamma}) = \sum_{y^n\tilde{\boldsymbol{\gamma}}\gamma\in\mathcal{L}(\mathcal{T}_{y^n})} a(y^n\tilde{\boldsymbol{\gamma}}\gamma). \quad (24)$$

By Lemma 7, $y^n\tilde{\boldsymbol{\gamma}}\gamma \in \mathcal{L}(\mathcal{T}_{y^n})$ if and only if $y^n\pi(\tilde{\boldsymbol{\gamma}})\gamma \in \mathcal{L}(\mathcal{T}_{y^n})$ for all permutations π. And as $a(y^n\tilde{\boldsymbol{\gamma}}\gamma) = a(y^n\pi(\tilde{\boldsymbol{\gamma}})\gamma)$, we can compute (24) in polynomial time in κ.

Consider next the sum over all $y^n\boldsymbol{\gamma} \in \mathcal{L}(\mathcal{T}_{y^n})$ in (19). Using Lemma 8

$$\begin{aligned}\sum_{y^n\boldsymbol{\gamma}\in\mathcal{L}(\mathcal{T}_{y^n})} \sum_{k=n}^{n+l(\boldsymbol{\gamma})-1} &\mathbb{P}(Y^{n+l(\boldsymbol{\gamma})} = y^n\boldsymbol{\gamma})\times \\ &\times \mathbb{P}(S \le k|Y^{n+l(\boldsymbol{\gamma})} = y^n\boldsymbol{\gamma}) \\ &= \sum_{y^n\boldsymbol{\gamma}\in\mathcal{I}(\mathcal{T}_{y^n})} \mathbb{P}(Y^{n+l(\boldsymbol{\gamma})} = y^n\boldsymbol{\gamma})\times \\ &\quad \times \mathbb{P}(S \le n + l(\boldsymbol{\gamma})|Y^{n+l(\boldsymbol{\gamma})} = y^n\boldsymbol{\gamma}).\end{aligned}$$

Applying Lemma 7 as before, we conclude that the right-hand side can be computed in polynomial time in κ.

It remains to prove that α_{m+1} and d_{m+1} can be computed in polynomial time in κ from α_m and d_m. This follows from the same argument, as it suffices to compute the differences $b(\mathcal{T}_{\boldsymbol{y}^*}) - b(\boldsymbol{y}^*)$ and $a(\boldsymbol{y}^*) - a(\mathcal{T}_{\boldsymbol{y}^*})$ for all $\boldsymbol{y}^*$ maximizing $g(\boldsymbol{y},\mathcal{T})$.

Appendix II
Proof of Theorem 11

Fix some $\lambda \ge 0$. Let us write $J_\lambda(T)$ as $\mathbb{E}(c(Y^T))$ where

$$c(y^n) \triangleq E((n-S)^+|Y^n = y^n) + \lambda\mathbb{P}(S > n|Y^n = y^n).$$

We say that the $\{\mathcal{A}_n\}$ are *nested* if, for any $n \ge 1$ and $\gamma \in \mathcal{Y}$, we have that $y^n \in \mathcal{A}_n$ implies $y^n\gamma \in \mathcal{A}_{n+1}$. We show that (22) implies that the $\{\mathcal{A}_n\}$ are nested, and that this in turn implies that the one-step look-ahead stopping rule is optimal. The second part of the proof is well known in the theory of optimal stopping and is referred as the *monotone case* (see, e.g., Chow *et al.* [7, Ch. 3]). Here we provide an alternative proof that emphasizes the tree structure of s.t.'s.

$y^n \in \mathcal{A}_n$ if and only if $\mathbb{E}(c(Y^{n+1})|Y^n = y^n) \ge c(y^n)$. We now show that

$$\mathbb{E}(c(Y^{n+1})|Y^n) \ge c(Y^n) \iff \mathbb{P}(S \le n|Y^n) \ge \lambda\mathbb{P}(S = n+1|Y^n). \quad (25)$$

Since $\{(X_i, Y_i)\}_{i\ge1}$ are i.i.d., S is also a (randomized) s.t. with respect to $\{Y_i\}_{i\ge1}$ by Lemma 8. It follows that

$$\begin{aligned}c(Y^{n+1}) &= E((n+1-S)^+|Y^{n+1}) + \lambda\mathbb{P}(S > n+1|Y^{n+1}) \\ &= \sum_{k=1}^{n} \mathbb{P}(S \le k|Y^{n+1}) + \lambda\mathbb{P}(S > n+1|Y^{n+1}) \\ &= \sum_{k=1}^{n-1} \mathbb{P}(S \le k|Y^n) + \mathbb{P}(S \le n|Y^n) \\ &\quad + \lambda\mathbb{P}(S > n|Y^n) - \lambda\mathbb{P}(S = n+1|Y^{n+1}) \\ &= c(Y^n) + \mathbb{P}(S \le n|Y^n) - \lambda\mathbb{P}(S = n+1|Y^{n+1})\end{aligned}$$

from which one deduces (25).

Next, we prove that the $\{\mathcal{A}_n\}$ are nested. By (25) this is equivalent to showing that, whenever for some y^n

$$\mathbb{P}(S \le n|Y^n = y^n) \ge \lambda\mathbb{P}(S = n+1|Y^n = y^n) \quad (26)$$

we also have

$$\begin{aligned}\mathbb{P}(S \le n+1|Y^{n+1} = y^n\gamma) \\ \ge \lambda\mathbb{P}(S = n+2|Y^{n+1} = y^n\gamma)\end{aligned} \quad (27)$$

for any $\gamma \in \mathcal{Y}$. Suppose that (26) holds for some y^n. Using the fact that S is a s.t. with respect to the Y_i's (Lemma 8) together with the hypothesis of the theorem yields for any γ

$$\begin{aligned}\mathbb{P}(S \le n+1|Y^{n+1} = y^n\gamma) &- \lambda\mathbb{P}(S = n+2|Y^{n+1} = y^n\gamma) \\ &\ge \mathbb{P}(S \le n|Y^n = y^n) - \lambda(S = n+2|Y^{n+1} = y^n\gamma) \\ &\ge \lambda\Big(\mathbb{P}(S = n+1|Y^n = y^n) - \mathbb{P}(S = n+2|Y^{n+1} = y^n\gamma)\Big) \\ &\ge 0\end{aligned}$$

and therefore (27) holds. Hence , the $\{\mathcal{A}_n\}$ are nested.

Let $\mathcal{T}^*$ be the tree corresponding to T^*_λ. The final step is to show that if the $\{\mathcal{A}_n\}$ are nested then $\mathcal{T}^0(\lambda) = \mathcal{T}^*$. To that aim we show that $\mathcal{I}(\mathcal{T}^*) \subset \mathcal{I}(\mathcal{T}^0(\lambda))$ and $(\mathcal{I}(\mathcal{T}^*))^c \subset (\mathcal{I}(\mathcal{T}^0(\lambda)))^c$. Pick an arbitrary $\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^0)$. Using Lemma 2, we compare $J_\lambda(\boldsymbol{y})$ with $\sum_\gamma J_\lambda(\mathcal{T}^0_{\boldsymbol{y}\gamma}(\lambda))$. We distinguish two cases. First suppose that $\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^*)$, i.e., $J_\lambda(\boldsymbol{y}) > \sum_\gamma J_\lambda(\boldsymbol{y}\gamma)$. Then

$$J_\lambda(\boldsymbol{y}) > \sum_{\gamma\in\mathcal{Y}} J_\lambda(\boldsymbol{y}\gamma) \ge \sum_{\gamma\in\mathcal{Y}} J_\lambda\left(\mathcal{T}^0_{\boldsymbol{y}\gamma}(\lambda)\right)$$

and, hence, $\boldsymbol{y} \notin \mathcal{L}(\mathcal{T}^0(\lambda))$. But since the $\{\mathcal{A}_n\}$ are nested, no prefix of $\boldsymbol{y}$ can be an element of $\mathcal{L}(\mathcal{T}^0(\lambda))$ and hence $\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^0(\lambda))$.

Second, assume $\boldsymbol{y} \notin \mathcal{I}(\mathcal{T}^*)$. If $l(\boldsymbol{y}) = \kappa$, then clearly $\boldsymbol{y} \notin \mathcal{I}(\mathcal{T}^0(\lambda))$. If $l(\boldsymbol{y}) < \kappa$, then $J_\lambda(\boldsymbol{y}) \le \sum_\gamma J_\lambda(\boldsymbol{y}\gamma)$ and we now show by induction that this implies that $\mathcal{T}^0_{\boldsymbol{y}}(\lambda) = \{\boldsymbol{y}\}$. Note first that as the $\{\mathcal{A}_n\}$ are nested, we have for any $\tilde{\boldsymbol{y}} \in \mathcal{I}(\mathcal{T}^0_{\boldsymbol{y}})$ (i.e., for any $\tilde{\boldsymbol{y}}$ with prefix $\boldsymbol{y}$)

$$J_\lambda(\tilde{\boldsymbol{y}}) \le \sum_{\gamma\in\mathcal{Y}} J_\lambda(\tilde{\boldsymbol{y}}\gamma). \quad (28)$$

Assume first that $\mathcal{T}^0_{\tilde{\boldsymbol{y}}}$ has depth one. Then (28) implies by Lemma 2 that $\mathcal{T}^0_{\tilde{\boldsymbol{y}}}(\lambda) = \{\tilde{\boldsymbol{y}}\}$. Suppose then that this is true for

all $\mathcal{T}^0_{\tilde{\boldsymbol{y}}}$ of depth at most $k-1$. Let $\mathcal{T}^0_{\tilde{\boldsymbol{y}}}$ have depth k. Then by the induction hypothesis and (28)

$$\sum_{\gamma\in\mathcal{Y}} J_\lambda(\mathcal{T}^0_{\tilde{\boldsymbol{y}}\gamma}(\lambda)) = \sum_{\gamma\in\mathcal{Y}} J_\lambda(\tilde{\boldsymbol{y}}\gamma) \geq J_\lambda(\tilde{\boldsymbol{y}})$$

and thus $\mathcal{T}^0_{\tilde{\boldsymbol{y}}}(\lambda) = \{\tilde{\boldsymbol{y}}\}$ by Lemma 2, concluding the induction step. This implies $\boldsymbol{y} \notin \mathcal{I}(\mathcal{T}^0(\lambda))$.

ACKNOWLEDGMENT

The authors are indebted to Prof. G. Wornell for providing them with generous funding support without which this work would not have been possible. We would also like to thank the anonymous reviewer for helpful comments.

REFERENCES

[1] F. J. Anscombe, H. J. Godwin, and R. L. Plackett, *Methods of Deferred Sentencing in Testing the Fraction Defective of a Continuous Output*, Supplement to the *J. Roy. Statist. Soc.*, vol. 9, no. 2, pp. 198–217, 1947.

[2] J. R. Baxter and R. V. Chacon, "Compactness of s.t.s," *Probability Theory and Related Fields*, vol. 40, no. 3, pp. 169–181, 1977.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. London, U.K.: Chapman & Hall, 1984.

[5] M. V. Burnashev, "Data transmission over a discrete channel with feedback: Random transmission time," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 250–265, 1976.

[6] M. V. Burnashev and H. Yamamoto, "On zero-rate error exponent for BSC with noisy feedback," *Probl. Inform. Transm.*, to be published.

[7] Y. S. Chow, R. Robbins, and D. Siegmund, *The Theory of Optimal Stopping*. New York: Dover, 1991.

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

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[10] R. L. Dobrushin, "Asymptotic bounds on the probability of error for the transmission of messages over a memoryless channel using feedback," *Probl. Kibern.*, vol. 8, pp. 161–168, 1963.

[11] G. A. Edgar, A. Millet, and L. Sucheston, "On compactness and optimality of stopping times," in *Martingale Theory in Harmonic Analysis and Banach Spaces*. New York: Springer-Verlag, 1982, pp. 36–61.

[12] M. Horstein, "Sequential transmission using noiseless Feedback," *IEEE Trans. Inf. Theory*, vol. IT-9, no. 3, pp. 136–143, 1963.

[13] Y.-H. Kim, A. Lapidoth, and T. Weissman, "On reliability of Gaussian channels with noisy feedback," in *Proc. 44th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sep. 2006.

[14] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2917–2929, Nov. 1998.

[15] S.-Y. R. Li, "Dynamic programming by exchangeability," *SIAM J. Comput.*, vol. 3, no. 3, pp. 463–472, 1989.

[16] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, 1971.

[17] G. V. Moustakides, "Optimal stopping times for detecting changes in distribution," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, 1986.

[18] M. Pollak, "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 14, no. 1, pp. 206–227, 1985.

[19] A. N. Shiryaev, *Optimal Stopping Rules*. New York: Springer-Verlag, 1978.

[20] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theor. Probl. and its Apploc.*, vol. 8, no. 1, pp. 22–46, 1963.

[21] A. G. Tartakovsky and V. V. Veeravalli, "General asymptotic Bayesian theory of qickest change detection," *Theor. Probl. and its Applic.*, vol. 49, no. 3, pp. 458–497, 2005.

[22] 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,*, ser. Lecture Notes, Monograph Series. Beachwood, OH: Instit. Math. Statist., 1994, vol. 23, pp. 346–358.

Urs Niesen (S'02) received the M.S. degree from the School of Computer and Communication Sciences at the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2005.

He is currently at the Massachusetts Institute of Technology (MIT), Cambridge, in the Department of Electrical Engineering and Computer Science, where he is working toward the Ph.D. degree. His research interests are in the area of communication and information theory.

Aslan Tchamkerten received the engineer physicist diploma in 2000 and the Ph.D. degree in communication in 2005, both from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

After graduation, he was a Postdoctoral Associate in the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge. In September 2008, he joined TELECOM ParisTech (ENST), Paris, France, as an Assistant Professor. His research interests are in information theory, statistical decision theory, and complexity.