Tracking Stopping Times

Urs Niesen, Aslan Tchamkerten, and Gregory Wornell

***Abstract*— Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of pairs of random variables, and let S be a bounded stopping time with respect to $\{X_i\}_{i=1}^{\infty}$. We propose the problem of finding a stopping time T with respect to $\{Y_i\}_{i=1}^{\infty}$ that optimally tracks S in the sense that T minimizes the average *reaction time* $\mathbb{E}(T-S)^+$ while keeping the *false-alarm probability* $\mathbb{P}(T<S)$ below a given threshold α.**

This problem has applications in many different areas. In this paper we present an application related to communication over a channel with noisy feedback.

I. Introduction

Consider a sensor that monitors the seismic activity of a volcano and sequentially sends the data to a remote analyzer. The analyzer, by observing only a noisy version of the collected data, has to raise an alarm as soon as an eruption is imminent. What is the loss in "forecast precision" incurred because of the noise? If the noisy data the analyzer receives is almost independent of the one obtained by the sensors, the analyzer will raise spurious or late alarms most of the time. On the other hand, if the analyzer receives the same data as the sensor, "optimal forecasting" may be possible.

The above situation provides a motivation for the following tracking stopping time problem. Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be an arbitrary sequence of pairs of random variables, where the (X_i, Y_i)'s take values in some finite alphabet $\mathcal{X} \times \mathcal{Y}$. Suppose that Alice observes $\{X_i\}_{i=1}^{\infty}$ and that she chooses a stopping time S with respect to $\{X_i\}_{i=1}^{\infty}$.[1] Having access only to $\{Y_i\}_{i=1}^{\infty}$, what is the best stopping time T Bob can find in order to minimize the expected delay $\mathbb{E}(T-S)^+$ while ensuring the probability of false-alarm $\mathbb{P}(T<S)$ to be below a certain threshold $\alpha \in [0,1]$?[2] We assume that Bob knows the distribution of $\{(X_i, Y_i)\}_{i=1}^{\infty}$ and the stopping rule S – but not the realizations of S. In the language of the above example, Alice and Bob represent the sensor and the analyzer, respectively, S the optimal time to raise an alarm given perfect observations, and T the optimal time to raise an alarm given noisy observations.

Another example where a stopping time needs to be tracked arises in the context of communication with feedback. It is well known that the presence of a noiseless feedback link allows to increase reliability given a certain communication delay (see, e.g., [1]). However, to take advantage of feedback, variable length codes are often necessary. The archetype example is communication with feedback over a binary erasure channel with erasure probability $\varepsilon \in (0,1)$ of a 1-bit message $m \in \{0,1\}$ (see, e.g., [2, Prob. 2.10]). On the one hand, any block coding strategy yields a strictly positive error probability. On the other hand, the variable length strategy of sending the same bit until a non-erasure occurs yields error free communication at a rate equal to capacity. Now suppose the feedback link is a binary erasure channel with crossover probability $p \in (0,1)$. Because of the noise in the feedback link, the first non-erased output symbol may not be recognized as such by the encoder. Hence the problem of synchronizing the transmitter and the receiver arises from the encoder observing through feedback only a noisy version of the symbols received by the decoder.

Instead of treating the synchronization issue resulting from the noisy feedback channel, let us consider the simpler problem of finding the minimum delay needed by the encoder to realize that the decoder has made a decision. In terms of our generic problem of tracking a stopping time, Alice and Bob represent now the decoder and the encoder, respectively. The sequence $\{X_i\}_{i=1}^{\infty}$ corresponds to the symbols fed into the feedback channel, whereas $\{Y_i\}_{i=1}^{\infty}$ corresponds to the output of the feedback channel. The stopping time S is the first time the decoder receives a non-erasure, and T the time the encoder stops retransmission. Here $\mathbb{E}(T-S)^+$ represents the delay it takes the encoder to realize that the decoder has made a decision, and we aim to minimize this delay given that the false-alarm probability $\mathbb{P}(T<S)$ is kept below a certain threshold α.

The tracking stopping time problem as defined above appears naturally in many different areas such as detection, forecasting, and communication. Somewhat surprisingly, and to the best of our knowledge, this problem does not appear in the literature.

This paper is organized as follows. In Section II, we formally define the problem of tracking a stopping time and provide an algorithmic solution to it. In Section III, we present two examples, one of which is the problem of communication over a binary erasure channel with noisy feedback described above. For each application, we derive a lower bound on the smallest reaction delay $\mathbb{E}(T-S)^+$ given that the false-alarm probability $\mathbb{P}(T<S)$ is kept below a certain threshold α. Section IV contains concluding remarks.

II. Problem Formulation

Given the sequences $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$, a stopping time S with respect to $\{X_i\}_{i=1}^{\infty}$, and the false-alarm level

This work was supported in part by NSF under Grant No. CCF-0515122, and by Draper Laboratory.

The authors are with the Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA 02139, USA. {uniesen,tcham,gww}@mit.edu

[1] An integer-valued random variable S is called a stopping time with respect to a sequence of random variables $\{X_i\}_{i=1}^{\infty}$ if, conditioned on $\{X_i\}_{i=1}^{n}$, the event $\{S=n\}$ is independent of $\{X_i\}_{i=n+1}^{\infty}$ for all $n \geq 1$.

[2] We use x^+ to denote $\max\{0,x\}$.

$\alpha \in [0,1]$, we aim to find

$$\beta(\alpha) \triangleq \min_{T:\mathbb{P}(T<S)\leq\alpha} \mathbb{E}(T-S)^{+} \tag{1}$$

where the minimization is over all stopping times T with respect to $\{Y_i\}_{i=1}^{\infty}$. Throughout the paper, we assume that S is bounded, i.e., there exists $S_0 \in \mathbb{N}$ such that $\mathbb{P}(S \leq S_0) = 1$. Without loss of optimality, we restrict the T's to be bounded by S_0 as well.

Note that the minimization problem (1) is a convex optimization problem since $\{T : \mathbb{P}(T < S) \leq \alpha\}$ is convex and since $\mathbb{E}(T-S)^{+}$ is convex with respect to T.[3] Therefore, the Lagrangian formulation yields

$$\beta(\alpha) = \max_{\lambda\geq 0} \min_{T} \left(J_\lambda(T) - \lambda\alpha\right)$$

where $J_\lambda(T) \triangleq \mathbb{E}(T-S)^{+} + \lambda\mathbb{P}(T < S)$ (see, e.g., [3]).

In Sections II-A and II-B we compute $J_\lambda(T)$ for deterministic and non deterministic stopping times T, respectively.

A. Deterministic Stopping Times

An element in $\mathcal{Y}^*$ will be denoted either by $\mathbf{y}$ or by y^k, depending on whether we want to emphasize the length of the sequence or not.[4] A stopping time is said to be deterministic if $\mathbb{P}(T = k|Y^k = y^k) \in \{0,1\}$ for all $y^k \in \mathcal{Y}^*$ and $k \geq 1$. To any deterministic stopping time T, we associate a unique $|\mathcal{Y}|$-ary tree[5] $\mathcal{T}$ having each node specified by some $y^k \in \mathcal{Y}^*$, where ρy^k represents the vertex path from the root ρ to the node. The depth of a node $y^k \in \mathcal{T}$ is defined as $l(y^k) \triangleq k$. The tree consisting only of the root is the trivial tree. A node $y^k \in \mathcal{T}$ is a leaf if $\mathbb{P}(T = k|Y^k = y^k) = 1$. We denote by $\mathcal{L}(\mathcal{T})$ the leaves of $\mathcal{T}$ and by $\mathcal{I}(\mathcal{T})$ the intermediate (or non-terminal) nodes of $\mathcal{T}$. The notation $T(\mathcal{T})$ is used to denote the stopping time 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 $\boldsymbol{y}$. The next example illustrates these notations.

Example 1. Let $\mathcal{Y} = \{0,1\}$ and $S_0 = 2$. The tree $\mathcal{T}$ depicted in Figure 1 corresponds to the deterministic stopping time T taking value one if $Y_0 = 1$ and value 2 if $Y_0 = 0$. 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 $\{0, 00, 01\}$. The subtree $\mathcal{T}_\rho$ is the same as $\mathcal{T}$. $\diamondsuit$

For a given stopping rule S, we now describe an algorithm constructing a sequence of stopping times $\{T(\mathcal{T}^m)\}_{m=0}^{M}$ 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 $\mathcal{T}^M \subset \mathcal{T}^{M-1} \subset \ldots \subset \mathcal{T}^0$ and $0 = \lambda_M \leq \lambda_{M-1} \leq \ldots \leq \lambda_1 \leq \lambda_0 = \infty$.[6] Second, for any $m \in \{0,\ldots,M\}$ and $\lambda \in (\lambda_m, \lambda_{m-1}]$ the tree $\mathcal{T}^{m-1}$ minimizes $J_\lambda(\mathcal{T}) \triangleq J_\lambda(T(\mathcal{T}))$ among all deterministic stopping times.

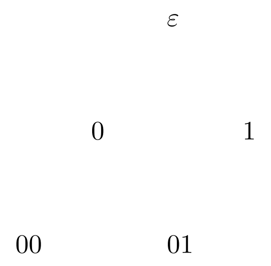

Fig. 1. The tree $\mathcal{T}$ corresponding to a stopping time bounded by $S_0 = 2$.

The algorithm builds upon ideas from the CART algorithm for the construction of classification and regression trees [4]. The remainder of this section adapts the arguments in [4, Chapter 10] to the problem at hand.

Given a stopping time T represented by its $|\mathcal{Y}|$-ary tree $\mathcal{T}$, we expand $J_\lambda(\mathcal{T})$ as

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

where $b(\boldsymbol{y}) \triangleq \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})\mathbb{E}\big((l(\boldsymbol{y}) - S)^{+}|\boldsymbol{Y} = \boldsymbol{y}\big)$, $a(\boldsymbol{y}) \triangleq \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})\mathbb{P}(S > l(\boldsymbol{y})|\boldsymbol{Y} = \boldsymbol{y})$, and $J_\lambda(\boldsymbol{y}) \triangleq b(\boldsymbol{y}) + \lambda a(\boldsymbol{y})$. We extend the definition of $J_\lambda(\cdot)$ to subtrees of $\mathcal{T}$ by setting $J_\lambda(\mathcal{T}_{\boldsymbol{y}}) \triangleq \sum_{\gamma\in\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})} J_\lambda(\gamma)$. With this definition we have

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

For a given $\lambda \geq 0$ and $\mathcal{T}$ define (if it exists) $\mathcal{T}(\lambda) \subset \mathcal{T}$ to be the subtree of $\mathcal{T}$ such that $J_\lambda(\mathcal{T}(\lambda)) \leq J_\lambda(\mathcal{T}')$ for all subtrees $\mathcal{T}' \subset \mathcal{T}$, and such that $\mathcal{T}(\lambda) \subset \mathcal{T}'$ for all subtrees $\mathcal{T}' \subset \mathcal{T}$ satisfying $J_\lambda(\mathcal{T}(\lambda)) = J_\lambda(\mathcal{T}')$. In words, among all subtrees yielding a minimal cost for a given λ, the tree $\mathcal{T}(\lambda)$ is the smallest. In the sequel $\mathcal{T}(\lambda)$ will be said to be optimal with respect to λ and $\mathcal{T}$.
Remark: Note that $\mathcal{T}_{\boldsymbol{y}}(\lambda)$ is different from $(\mathcal{T}(\lambda))_{\boldsymbol{y}}$. Indeed, $\mathcal{T}_{\boldsymbol{y}}(\lambda)$ refers to the optimal subtree of $\mathcal{T}_y$ with respect to λ, whereas $(\mathcal{T}(\lambda))_{\boldsymbol{y}}$ refers to subtree rooted in $\boldsymbol{y}$ of the optimal tree $\mathcal{T}(\lambda)$.

Given a $|\mathcal{Y}|$-ary tree $\mathcal{T}$ and a $\lambda \geq 0$, the following lemma shows that $\mathcal{T}(\lambda)$ always exists and characterizes $\mathcal{T}(\lambda)$ and $J_\lambda(\mathcal{T}(\lambda))$ using dynamic programming.

Lemma 1. *Given a $|\mathcal{Y}|$-ary tree $\mathcal{T}$ and $\lambda \geq 0$, starting with the root node recursively compute*

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

[3] Given $\delta \in [0,1]$, the convex combination of two stopping times T_1 and T_2, denoted by $T = \delta T_1 + (1-\delta)T_2$, is obtained as follows. Let U be a random variable uniformly distributed within the interval $[0,1]$, and independent of T_1 and T_2. By definition, the stopping time T equals to T_1 if $U \leq \delta$ and equals to U_2 if $U > \delta$.

[4] The set $\mathcal{Y}^*$ represents the set of all finite sequences over $\mathcal{Y}$.

[5] A tree is $|\mathcal{Y}|$-ary if all its nodes have either zero or exactly $|\mathcal{Y}|$ children

[6] The symbol $\subset$ denotes inclusion, not necessarily strict.

and recursively construct

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

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 $T(\lambda)$ will stop. Indeed, Lemma 2 states that the tree corresponding to the optimal stopping time of a smaller λ is nested into the tree corresponding to the optimal stopping time of a larger λ. In other words, if $\lambda\le\tilde{\lambda}$, in order to find $\mathcal{T}(\lambda)$ we can restrict our search to subtrees of $\mathcal{T}(\tilde{\lambda})$.

Lemma 2. *Given a tree $\mathcal{T}$, if $\lambda\le\tilde{\lambda}$ then $\mathcal{T}(\lambda)\subset\mathcal{T}(\tilde{\lambda})$.*

Proof: Note first that $\sum_{\gamma\in\mathcal{Y}}a(\boldsymbol{y}\gamma)\le a(\boldsymbol{y})$ and $\sum_{\gamma\in\mathcal{Y}}b(\boldsymbol{y}\gamma)\ge b(\boldsymbol{y})$. Hence if $\mathcal{T}(\tilde{\lambda})=\{\rho\}$ then $\mathcal{T}(\lambda)=\{\rho\}$. The result follows now by induction on the depth of the tree and using Lemma 1. ■

The next theorem represents a key result. Given a tree $\mathcal{T}$, this theorem characterizes the smallest value λ can take for which $\mathcal{T}(\lambda)=\mathcal{T}$. In the sequel we use $a(\mathcal{T}_{\boldsymbol{y}})$ to denote $\sum_{\gamma\in\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})}a(\gamma)$ and $b(\mathcal{T}_{\boldsymbol{y}})$ to denote $\sum_{\gamma\in\mathcal{L}(\mathcal{T}_{\boldsymbol{y}})}b(\gamma)$.

Theorem 3. *For a non trivial tree $\mathcal{T}$, define for any intermediate node $\boldsymbol{y}\in\mathcal{I}(\mathcal{T})$*

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

We have

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

Proof: Let $\mathcal{T}$ be a non trivial tree. We have

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

An easy computation reveals that $a(\mathcal{T}_{\boldsymbol{y}})\le a(\boldsymbol{y})$, hence the following implications hold:

$$\begin{aligned}g(\boldsymbol{y},\mathcal{T})\le\lambda&\Longleftrightarrow J_\lambda(\boldsymbol{y})\ge J_\lambda(\mathcal{T}_{\boldsymbol{y}})\\ g(\boldsymbol{y},\mathcal{T})<\lambda&\Longleftrightarrow J_\lambda(\boldsymbol{y})>J_\lambda(\mathcal{T}_{\boldsymbol{y}})\ .\end{aligned}$$

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

$$J_\lambda(\boldsymbol{y})>J_\lambda(\mathcal{T}_{\boldsymbol{y}})\tag{2}$$

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

We first show by induction that if

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

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

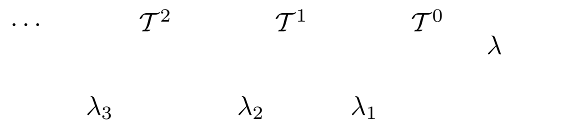

Fig. 2. For all $i\in\{0,1,\ldots,M-1\}$ the tree $\mathcal{T}^i$ is the smallest tree minimizing the cost $J_\lambda(\cdot)$ for any $\lambda\in(\lambda_{i+1},\lambda_i]$.

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

and since $J_\lambda(\mathcal{T}_{\boldsymbol{y}})<J_\lambda(\boldsymbol{y})$ by (2), we have $\mathcal{T}_{\boldsymbol{y}}(\lambda)=\mathcal{T}_{\boldsymbol{y}}$ by Lemma 1, 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

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

In this case there exists $\boldsymbol{y}\in\mathcal{I}\big(\mathcal{T}\big)$ such that $J_\lambda\big(\mathcal{T}_{\boldsymbol{y}}\big)=J_\lambda(\boldsymbol{y})$. 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

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

and thus $\mathcal{T}(\lambda)\neq\mathcal{T}$ by Lemma 1. 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 1.

Finally, when $\lambda<\max_{\boldsymbol{y}\in\mathcal{I}(\mathcal{T})}g\big(\boldsymbol{y},\mathcal{T}\big)$ then $\mathcal{T}(\lambda)\neq\mathcal{T}$ follows from the previous case and Lemma 2. ■

Let $\mathcal{T}^0$ denote the full tree of depth S_0. Starting with $\lambda_0=\infty$, for $m=1,2,\ldots$ recursively define

$$\lambda_m\triangleq\inf\{\lambda\le\lambda_{m-1}:\mathcal{T}^{m-1}(\lambda)=\mathcal{T}^{m-1}\ \}$$

with $\lambda_1\triangleq\infty$ if the set over which the infimum is taken is empty. Hence, for two consecutive transition points λ_m and λ_{m+1} we have $\mathcal{T}(\lambda)=\mathcal{T}(\lambda_m)$ for all $\lambda\in(\lambda_{m+1},\lambda_m]$ as shown in Figure 2.

The following corollary is a consequence of Lemma 2 and Theorem 3.

Corollary 4. *For $m=1,2,\ldots$*

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

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

where $\mathcal{D}(\mathcal{T},\boldsymbol{y})\triangleq\mathcal{T}_{\boldsymbol{y}}\setminus\{\boldsymbol{y}\}$ denotes the descendants of $\boldsymbol{y}$ in $\mathcal{T}$ and where $\mathcal{T}^0$ is the full tree of depth S_0. The above iteration stops at $m=M$ with $\mathcal{T}^m=\{\rho\}$.

B. Randomized Stopping Times

So far we have imposed T to be a deterministic stopping time. We now remove this restriction and allow T to be randomized. A randomized stopping time is such that to each node $\boldsymbol{y} \in \widetilde{\mathcal{T}}$ is associated a weight given by the stopping probability $\mathbb{P}(T = k | Y^k = y^k, T \geq k) \in [0, 1]$. The leaf nodes have weights equal to 1, whereas the non-terminal nodes have weights in $[0, 1)$. This generalizes the deterministic stopping times where the leaf nodes have weight equal to 1 and the intermediate nodes weights equal to 0.

One can show the following lemma (see, e.g., [5]):

Lemma 5. *Any bounded randomized stopping time $\widetilde{T}$ can be written as a convex combination of bounded deterministic stopping times.* □

Theorem 6. *The function $\beta(\alpha)$ is convex and piecewise linear with the set of break-points $\{(\alpha_m, \beta_m)\}_{m=0}^{M}$ given by $\alpha_m = \mathbb{P}(T(\mathcal{T}^m) < S)$ and $\beta_m = \mathbb{E}(T(\mathcal{T}^m) - S)^+$.*[7]

Proof: From Lemma 6, the cost $J_\lambda(\widetilde{\mathcal{T}})$ of a randomized stopping time $\widetilde{T}$ can be written as the convex combination of costs of non-randomized stopping times, i.e.,

$$J_\lambda(\widetilde{\mathcal{T}}) = \sum_{\mathcal{T}} p_{\mathcal{T}} J_\lambda(\mathcal{T}) \tag{3}$$

with $p_{\mathcal{T}} \geq 0$ and $\sum_{\mathcal{T}} p_{\mathcal{T}} = 1$. From (3) one deduces that the set of achievable pairs $\left(\mathbb{P}(T < S), \mathbb{E}(T - S)^+\right)$ is convex. Consider now two consecutive pairs of achievable points $(\alpha_{m-1}, \beta_{m-1})$ and (α_m, β_m). On the one hand, the segment connecting these two points is achievable by convexity. On the other hand, this segment has slope $-\lambda_m$ since $J_{\lambda_m}(\mathcal{T}^m) = J_{\lambda_m}(\mathcal{T}^{m-1})$. Suppose there exists a stopping time $\widetilde{\mathcal{T}}$ yielding a point (α, β) below the line carried by that segment. Then $J_{\lambda_m}(\widetilde{\mathcal{T}}) < J_{\lambda_m}(\mathcal{T}^m)$, a contradiction by (3) and the definition of $\mathcal{T}^m$. ■

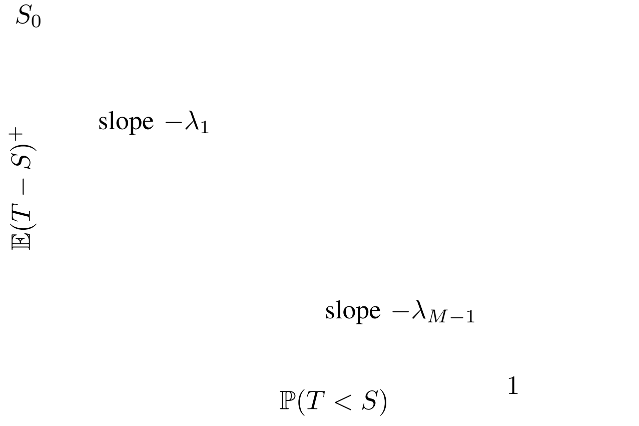

Fig. 3. Optimal operating points in the $\left(\mathbb{P}(T < S), \mathbb{E}(T - S)^+\right)$ plane.

Figure 3 illustrates a typical shape of $\beta(\alpha)$. Any point on $\beta(\alpha)$ that is not a break point can be achieved by a randomized stopping time given by the convex combination of two deterministic stopping times $\mathcal{T}^m$ and $\mathcal{T}^{m+1}$. From Corollary 4 and Theorem 6 we deduce Algorithm 1 below that fully characterizes $\beta(\alpha)$ by computing its break-points.

Algorithm 1 Compute the break-points $\{\alpha_m, \beta_m\}_{m=0}^{M}$ of $\beta(\alpha)$.

```
$\mathcal{T}^0 \Leftarrow$ full tree of depth $S_0$
$\lambda_0 \Leftarrow \infty$
$m \Leftarrow 0$
repeat
    $m \Leftarrow m + 1$
    $\lambda_m \Leftarrow \max_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^{m-1})} g(\boldsymbol{y}, \mathcal{T}^{m-1})$
    $\mathcal{T}^m \Leftarrow \mathcal{T}^{m-1} \setminus \bigcup_{\boldsymbol{y} \in \mathcal{I}(\mathcal{T}^{m-1}):\, g(\boldsymbol{y}, \mathcal{T}^{m-1}) = \lambda_m} \mathcal{D}(\mathcal{T}^{m-1}, \boldsymbol{y})$
    $\alpha_m \Leftarrow \mathbb{P}(T(\mathcal{T}^m) < S)$
    $\beta_m \Leftarrow \mathbb{E}(T(\mathcal{T}^m) - S)^+$
until $\mathcal{T}^m = \{\rho\}$
```

An analytical expression for $\beta(\alpha)$ is hard to obtain in general. Nevertheless, from Theorem 6 one can lower bound $\beta(\alpha)$ as

$$\beta(\alpha) \geq \beta(0) + \alpha\beta'(0+) \tag{4}$$

where $\beta'(0+)$ denotes the right derivative of β at $\alpha = 0$ (which exists by piecewise linearity of $\beta(\alpha)$). By Corollary 4, if $\lambda_1 < \infty$ then $\beta(0)$ is achieved by the full tree $\mathcal{T}^0$, and, if $\lambda_1 = \infty$, $\beta(0)$ is achieved by $\mathcal{T}^1$ which is a strict subtree of $\mathcal{T}^0$. Hence (4) can be written as

$$\beta(\alpha) \geq \beta(0) - \alpha \min\{\lambda_1, \lambda_2\}. \tag{5}$$

III. Examples

In this section, we consider two examples and, applying (5), we get simple analytical bounds on $\beta(\alpha)$.

Example 2. Let $\{X_i\}_{i=1}^{\infty}$ be i.i.d. Bernoulli$(\frac{1}{2})$ and let $\{Y_i\}_{i=1}^{\infty}$ represent the observation of the X_i's through a binary symmetric channel with crossover probability $p \in (0, \frac{1}{2})$. Consider the stopping time S defined as

$$S \triangleq \begin{cases} 1 & \text{if } X_1 = 1 \\ S_0 & \text{else.} \end{cases}$$

For $S_0 = 2$, the tree corresponding to this (deterministic) stopping time is depicted in Figure 1.

Since $p \in (0, \frac{1}{2})$, it is clear that whenever $\mathcal{T}$ is not the full tree of depth S_0, we have $\mathbb{P}(T(\mathcal{T}) < S) > 0$, hence

$$\beta(0) = \mathbb{E}(T(\mathcal{T}^0) - S)^+ = \frac{1}{2}(S_0 - 1).$$

From Corollary 4 an easy computation yields

$$\lambda_1 = \frac{1-p}{p}(S_0 - 1),$$

[7] $\beta(\alpha)$ is defined in (1).

and, using (5), we get

$$\beta(\alpha) \geq (S_0 - 1)\Big(\frac{1}{2} - \alpha\frac{1-p}{p}\Big). \tag{6}$$

Let us comment on (6). Consider any two correlated sequences $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ and a stopping time S with respect to the X_i's. Intuition tells us that there are two factors that affect $\beta(\alpha)$. The first is the correlation between the X_i's and Y_i's, in the above example given by the crossover probability p of the observation channel. The lower the correlation, the higher $\beta(\alpha)$ will be. 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 S_0, since with probability $1/2$ a leaf of depth 1 is reached, and with probability $1/2$ a leaf with depth S_0 is attained. ♢

Example 3. We consider 1-bit message feedback communication when the forward and the feedback channels are binary erasure channels with erasure probabilities ε and p, respectively. We refer the reader to Section I for the general problem setting. We use the following simple transmission scheme (which is optimal in the case of noiseless feedback). The decoder keeps sending 0 over the feedback channel until time S, the first time a non-erasure occurs or S_0 time units have elapsed. From that point on, the decoder sends 1. The encoder keeps sending the message bit it wants to deliver until time T, the first time it receives a 1 from the feedback channel or S_0 time units have elapsed. S_0 plays here the role of a "time-out".

Let us focus on $\beta(\alpha)$. One can show that $\lambda_1 = \infty$ and therefore the bound (5) becomes $\beta(\alpha) \geq \beta(0) - \alpha\lambda_2$, where $\lambda_2 = \max_{\boldsymbol{y}\in\mathcal{I}(\mathcal{T}^1)} g\big(\boldsymbol{y}, \mathcal{T}^1\big)$ from Corollary 4. A somewhat involved computation yields

$$\beta(\alpha) \geq \left(\frac{p}{1-p} - \varepsilon^{1-S_0}\alpha\right)(1 + o(1)) \tag{7}$$

as $S_0 \to \infty$.

The delay $\beta(\alpha)$ is interpreted as the time it takes the encoder to realize that the decoder has made a decision. Equation (7) relates this delay to the channel parameters ε and p, the probability α of stopping retransmission too early, and the value of the "time-out" S_0. For the communication scheme considered here, there are two events leading to decoding errors. The event $\{X_{S_0} = 0\}$, indicating that only erasures were received by the decoder until time S_0, 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 overall probability of error $\mathbb{P}(\mathcal{E})$ can be bounded as

$$\max\{\alpha, \varepsilon^{S_0}\} \leq 2\mathbb{P}(\mathcal{E}) \leq \alpha + \varepsilon^{S_0}.$$

It is then reasonable to choose $S_0 = \frac{\log\alpha}{\log\varepsilon}$, i.e., to scale S_0 with α so that both sources of errors have the same weight. This results in a delay of

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

as $\alpha \to 0$.

Now suppose that the communication rate R is computed with respect to the delay from the time communication starts until the time the encoder realizes that the decoder has made a decision, i.e., $\mathbb{E}S + \mathbb{E}(T-S)^+ = \mathbb{E}(\max\{S,T\})$. We conclude that the "send until a non-erasure" strategy asymptotically achieves a rate that is upper bounded as

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

Note that when $\varepsilon < p/(1-p)$, our 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 $1-\varepsilon$. ♢

IV. Conclusion

We introduced the tracking stopping time problem. This problem asks for the "closest" stopping time T to a given stopping time S based on noisy observations of the data over which S is defined. We provided a solution for the case of bounded stopping times defined over discrete time processes taking values in a finite alphabet.

Acknowledgement

The authors would like to thank Prof. Robert Gallager and Baris Nakiboglu for helpful discussions.

References

[1] M. Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, 9:136–143, July 1963.

[2] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, Budapest, 1968.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] L. Breiman, J. H. Friedman, and R. Olshen. *Classification and regression trees*. Chapman & Hall, 1993.

[5] J.R. Baxter and R.V. Chacon. Compactness of stopping times. *Probability Theory and Related Fields*, 40(3):169–181, 1977.