Abstract—Consider a sequence of independent random variables that undergo a transient change in distribution from $P_0$ to $P_1$ at an unknown instant. The sequence is sequentially observed under a sampling constraint. This paper addresses the question of finding the minimum sampling rate needed in order to detect the change “as efficiently” as under full sampling.

The problem is cast into a Bayesian setup where the change, assumed to be of fixed known duration $n$, occurs randomly and uniformly within a time frame of size $A_n = 2^{\omega(n)}$ for some known uncertainty parameter $\omega > 0$. It is shown that, for any fixed $\alpha \in (0, D(P_1 \| P_0))$, as long as the sampling rate is of order $\omega(1/n)$ the change can be detected as quickly as under full sampling, in the limit of vanishing false-alarm probability. The delay, in this case, is some linear function of $n$. Conversely, if $\alpha > D(P_1 \| P_0)$ or if the sampling rate is $o(1/n)$ reliable detection is impossible—the false-alarm probability is bounded away from zero or the delay is $\Theta(2^{\omega(n)})$.

This paper illustrates this result through a recently proposed asynchronous communication framework. Here, the receiver observes mostly pure background noise except for a brief period of time, starting at an unknown instant, when data is transmitted thereby inducing a local change in distribution at the receiver. For this model, capacity per unit cost (minimum energy to transmit one bit of information) and communication delay were characterized and shown to be unaffected by a sparse sampling at the receiver as long as the sampling rate is a non-zero constant. This paper strengthens this result and shows that it continues to hold even if the sampling rate tends to zero at a rate no faster than $\omega(1/B)$, where $B$ denotes the number of transmitted message bits. Conversely, if the sampling rate decreases as $o(1/B)$, reliable communication is impossible.

Index Terms—Asynchronous communication; bursty communication; capacity per unit cost; energy; change detection; hypothesis testing; sequential analysis; sensor networks; sparse communication; sampling; synchronization; transient change

I. INTRODUCTION

Asynchrony and asynchronous communication is investigated when data transmission occurs very infrequently, at random moments. For such a setting characterized information theoretic limits in terms of transmitter/receiver energy consumption and communication delay for discrete-time systems. Energy consumption at the transmitter is modeled in the usual way by assigning a cost to each channel input. Energy consumption at the receiver is captured by the channel output sampling rate. This is motivated by the fact that in practice one of the receiver’s most power consuming functions is the sampling rate of the analog-to-digital converter.

Asynchronism in [1] is caused by the transmitter’s source of information, which is assumed to be bursty. Decoding happens on a sequential basis and communication (detection) delay is naturally defined as the elapsed time between the instant when data starts being sent and the instant when it is decoded. In turn, sampling rate is defined as the typical relative number of channel outputs observed until decoding happens. Specifically, sampling rate $\rho$ is said to be achievable if

$$\frac{\text{number of output samples taken until time } \tau}{\tau} \leq \rho$$

with high probability, where $\tau$ denotes the random decoding time.

The main result in [1] states that a sparse output sampling impacts neither capacity per unit cost nor communication delay asymptotically: for any $\rho > 0$ capacity per unit cost and communication delay under sampling constraint $\rho$ are (asymptotically) the same as for $\rho = 1$.

In fact, a stronger result holds as we show in this paper. Capacity per unit cost and communication delay are unaffected even if the sampling rate tends to zero at

$$\rho = \omega(1/B)$$

the error probability of any coding scheme is bounded away from zero.

At the heart of the achievability result is a generic sequential procedure for detecting a transient change in distribution of fixed known duration over a time series. Before and after the change observations are supposed to be drawn from a common nominal product distribution, whereas during the change observations are supposed to be distributed according to some “change” product distribution. The detection procedure has the following properties in the limit of vanishing probability of false-alarm:

1. it detects changes of minimal duration,
2. it detects changes with minimal delay,
3. it minimizes sampling rate.

Property 1. means that no procedure can detect changes of shorter duration, irrespectively of its delay and sampling rate.

1 Throughout the paper we use the standard “big-O” Landau notation to characterize growth rates (see, e.g., Chapter 3). These growth rates, e.g., $o(1)$, are intended in the limit $B \to \infty$, unless stated otherwise.
Property 2. and 3. mean that among all procedures that detect transient changes, the proposed procedure simultaneously minimizes delay and sampling rate.

Related works

This work addresses the problem of communicating asynchronously at minimum energy using a sequential analysis framework. Accordingly, we review related works in communication and sequential analysis.

COMMUNICATION: The basic asynchronous communication setup investigated in this paper was originally proposed in [3] and developed in a series of works [4]–[10]. These works investigate capacity (or capacity per unit cost) under various notions of delay; codeword length, typical delay, expected delay. In all these works, the communication regime of interest, dubbed “strong asynchronism,” is such that data transmission is very bursty. Data transmission typically happens on a time horizon which is exponential in the size of the message—the subexponential regime is of limited theoretical interest as it reduces to the synchronous setting [5].

In the exponential asynchronism regime channel outputs mostly correspond to pure noise. Hence a natural question is whether the receiver needs to constantly be in the “listening” mode and observe (sample) all channel outputs. Surprisingly, full output sampling is not necessary. In [1] it is shown that capacity per unit cost and minimum typical delay, established in [8] under full sampling, are (asymptotically) unaffected by a sparse output sampling as long as the sampling rate $\rho$ is non-zero. For this result to hold, it is important to allow adaptive sampling. In fact, under non-adaptive sampling capacity per unit cost is achievable but delay is impacted by a multiplicative factor of $1/\rho$.

SEQUENTIAL ANALYSIS: Decoding in the above communication setup can be cast into a change-detection sequential analysis framework. Specifically, decoding amounts to detecting and isolating the cause of a transient change in distribution under a sampling constraint. The correspondences between the two problems are the following:

- nominal distribution $\leftrightarrow$ pure noise
- change duration $\leftrightarrow$ codeword length
- posterior distributions $\leftrightarrow$ set of output distributions induced by the codewords

Before data transmission, the receiver observes pure noise generated by some known nominal distribution. Once data transmission starts, the distribution changes according to the sent codeword. Decoding happens based on a sampling strategy, a stopping rule defined on the sampled process, and an isolation rule which maps the stopped sampling process into one of the possible messages. Unlike classical sequential analysis frameworks, the model we consider includes coding which translates into the ability to optimize posterior distributions, say, to maximize rate. Rate, is specific to communication and ties the number of posterior hypothesis with detection delay or the change duration. By contrast, in sequential analysis the number of posterior hypothesis is typically fixed which, in the vanishing error probability regime, would correspond to a zero rate regime.

Detection and isolation for non-transient changes and without sampling constraint was investigated in [12]. The purely detection problem (single posterior distribution hence no isolation rule) of reacting to a transient change in distribution was investigated in a variety of works. In [13], [14, Chap. 3], for instance, the CUSUM detection procedure, originally proposed to detect non-transient changes, is investigated for detecting transient changes of given length. In [15] Section II.c], a variation of the CUSUM procedure is shown to achieve minimal detection delay in a certain asymptotic regime where the duration of the change is tied to a (vanishing) false-alarm probability constraint. In [16] the authors investigate a setup where the duration of the change is random and where the evolution of the pre-change, in-change, and post-change states is Markovian. Finally, [17] proposed another interesting variation of the CUSUM procedure that operates under a sampling constraint and that is tailored for detecting non-transient changes. This procedure has the salient feature of skipping samples in the event that a change is unlikely to have occurred. Optimality of this procedure in the minimax (non-Bayesian) setting for both transient and non-transient changes was recently established in [18].

Paper organization

Section II is devoted to the transient change-detection problem and Section III is devoted to the asynchronous communication model and considers the performance metrics proposed in [1]. Note that the main result in Section III relies crucially on the detection procedure proposed in Section II. Section IV is devoted to the proofs.

II. TRANSIENT CHANGE-DETECTION

A. Model

Let $P_0$ and $P_1$ be distributions defined over some finite alphabet $\mathcal{Y}$ and with finite divergence $\mathcal{D}$.

$$D(P_1||P_0) \overset{\text{def}}{=} \sum_y P_1(y) \log[P_1(y)/P_0(y)].$$

Let $\nu$ be uniformly distributed over

$$\{1, 2, \ldots, A_n = 2^n\}$$

where the integer $A_n$ denotes the uncertainty level and where $\alpha$ the corresponding uncertainty exponent, respectively.

Given $P_0$ and $P_1$ process $\{Y_t\}$ is defined as follows. Conditioned on the value of $\nu$, the $Y_t$’s are i.i.d. according to $P_0$ for

$$1 \leq t < \nu$$

or

$$\nu + n \leq t \leq A_n + n - 1$$

Throughout the paper logarithms are always intended to be to the base 2.
and i.i.d. according to $P_l$ for $\nu \leq t \leq \nu + n - 1$. Process $\{Y_t\}$ is thus i.i.d. $P_0$ except for a brief period of duration $n$ where it is i.i.d. $P_1$.

A statistician, with the knowledge of $n, \alpha, P_0,$ and $P_1$, observes $\{Y_t\}$ sequentially according to a sampling strategy:

**Definition 1** (Sampling strategy). A sampling strategy with respect to a stochastic process $\{Y_t\}$ consists of a collection of random time indices

$$S = \{S_1, \ldots, S_t\} \subseteq \{1, \ldots, A_n + n - 1\}$$

where $S_i < S_j$ for $i < j$. Time index $S_i$ is interpreted as the $i$th sampling time. Sampling may be adaptive which means that each sampling time can be a function of past observations. This means that $S_i$ is an arbitrary value in $\{1, \ldots, A_n + n - 1\}$, possibly random but independent of $Y_i^{A_n + n - 1}$, and for $j \geq 2$

$$S_j = g_j(\{Y_{S_i} : i < j\})$$

for some (possibly randomized) function $g_j : Y^{j-1} \to \{S_{j-1} + 1, \ldots, A_n + n - 1\}$.

The statistician wishes to detect the change in distribution by means of a sampling strategy $S$ and a stopping rule $\tau$ relative to the (natural filtration induced by the) sampled process $\{Y_{S_1}, Y_{S_2}, \ldots\}$. Since there are at most $A_n + n - 1$ sampling times, $\tau$ is bounded by $A_n + n - 1$.

A given pair $(S, \tau)$, from now on referred to as a “detector,” will be evaluated in terms of its probability of false-alarm, its detection delay, and its sampling rate.

**Definition 2** (False-alarm probability). For a given detector $(S, \tau)$ the probability of false-alarm is defined as

$$P(\tau < \nu) = P_0(\tau < \nu)$$

where $P_0$ denotes the $P_0$-product distribution.

**Definition 3** (Detection delay). For a given detector $(S, \tau)$ and $\epsilon > 0$, the delay, denoted by $d((S, \tau), \epsilon)$, is defined as the minimum $l \geq 0$ such that

$$P(\tau - \nu \leq l - 1) \geq 1 - \epsilon.$$

**Definition 4** (Sampling rate). For a given detector $(S, \tau)$ and $\epsilon > 0$, the sampling rate, denoted by $\rho((S, \tau), \epsilon)$, is defined as the “typical” relative number of samples taken until time $\tau$. Specifically, it is defined as the minimum $r \geq 0$ such that

$$P(|S_{\tau}|/\tau \leq r) \geq 1 - \epsilon$$

where

$$S_{\tau} \overset{\text{def}}{=} \{S \in S : S \leq \tau\} \overset{\text{(1)}}{=} \{S \in S : S \leq \tau\}$$

$^4$Notice that $\ell$, the total number of samples, may be random under adaptive sampling but also under non-adaptive sampling since the strategy may be randomized (but still independent of $Y_1^{A_n + n - 1}$).

$^5$Recall that a (deterministic or randomized) stopping time $\tau$ with respect to a sequence of random variables $Y_1, Y_2, \ldots$ is a positive, integer-valued, random variable such that the event $\{\tau = t\}$, conditioned on the realization of $Y_1, Y_2, \ldots, Y_t$, is independent of the realization of $Y_{t+1}, Y_{t+2}, \ldots$, for all $t \geq 1$.

denotes the number of samples taken up to time $t$.

**Definition 5** (Achievable sampling rate). Fix $\alpha \geq 0$ and fix a sequence of non-increasing values $\{\rho_n\}$ with $0 \leq \rho_n \leq 1$. Sampling rates $\{\rho_n\}$ are said to be achievable at uncertainty exponent $\alpha$ if there exists a sequence of detectors $\{(S_n, \tau_n)\}$ such that for all $n$ large enough

1) $(S_n, \tau_n)$ operates under uncertainty level $A_n = 2^{\alpha n}$,
2) the false-alarm probability is at most $\epsilon_n$,
3) the sampling rate satisfies $\rho((S_n, \tau_n), \epsilon_n) \leq \rho_n$,
4) the delay satisfies

$$\frac{1}{n} \log(d((S_n, \tau_n), \epsilon_n)) \leq \epsilon_n$$

for some sequence of nonnegative numbers $\{\epsilon_n\}$ such that $\epsilon_n \xrightarrow{n \to \infty} 0$.

Two comments are in order. First note that samples taken after time $\tau$ play no role in our performance metrics. Hence, from now on and without loss of generality we assume that the last sample is taken at time $\tau$, i.e., that the sampled process is truncated at time $\tau$. The truncated sampled process is thus given by the collection of sampling times $S_{\tau}$ (see (1)). In particular, we have

$$|S_{\tau}| \geq |S_t| \quad 1 \leq t \leq A_n + n - 1. \quad (2)$$

The second comment concerns the delay constraint 4). Without a delay constraint the problem is obviously trivial; by considering the trivial stopping time $\tau = A_n$ we achieve zero probability of false-alarm and sampling rate $1/A_n$. The delay constraint intends to capture the fact that the detector locates the change very accurately at the scale of the uncertainty level. However, we will see that this constraint is in fact not very restrictive; whenever detection is possible with subexponential delay detection is also possible with a linear delay.

**Notational convention**

We shall use $d_n$ and $\rho_n$ instead of $d((S_n, \tau_n), \epsilon_n)$ and $\rho((S_n, \tau_n), \epsilon_n)$, respectively, leaving out any explicit reference to $(S_n, \tau_n)$ and to the sequence of nonnegative numbers $\{\epsilon_n\}$ which we assume satisfies $\epsilon_n \to 0$.

**B. Results**

Define

$$n^*(\alpha) \overset{\text{def}}{=} \frac{n\alpha}{D(P_1||P_0)} = \Theta(n). \quad (3)$$

**Theorem 1** (Detection, full sampling). Under full sampling ($\rho_n = 1$):

1) the supremum of the set of achievable uncertainty exponents is $D(P_1||P_0)$;
2) any detector that achieves uncertainty exponent $\alpha \in (0, D(P_1||P_0))$ has a delay that satisfies

$$\liminf_{n \to \infty} \frac{d_n}{n^*(\alpha)} \geq 1;$$
3) any uncertainty exponent \( \alpha \in (0, D(P_1||P_0)) \) is achievable with delay satisfying
\[
\limsup_{n \to \infty} \frac{d_n}{n^{\ast}(\alpha)} \leq 1.
\]
Hence, the shortest detectable change is of size
\[
n_{\min}(A_n) = \frac{\log A_n}{D(P_1||P_0)}(1 \pm o(1)) \quad (4)
\]
by Claim 1) of Theorem\( \ref{thm:main} \) assuming \( A_n \gg 1 \). In this regime, change duration and minimum detection delay are essentially the same by Claims 2)-3) and \( \ref{eq:uncertainty} \), i.e.,
\[
n^{\ast}(\alpha = (\log A_n)/n_{\min}(A_n)) = n_{\min}(A_n)(1 \pm o(1))
\]
whereas in general minimum detection delay could be smaller than change duration.

The next theorem says that the minimum sampling rate needed to achieve the same detection delay as under full sampling decreases essentially as \( 1/n \):

**Theorem 2** (Detection, sparse sampling). Fix \( \alpha \in (0, D(P_1||P_0)) \). Any sampling rate
\[
\rho_n = \omega(1/n)
\]
is achievable with delay satisfying
\[
\limsup_{n \to \infty} \frac{d_n}{n^{\ast}(\alpha)} \leq 1.
\]
Conversely, if
\[
\rho_n = o(1/n)
\]
the detector samples only from distribution \( P_0 \) (i.e., it completely misses the change) with probability bounded away from zero, regardless of the delay. Moreover, \( d_n = \Theta(A_n = 2^{\alpha n}) \) for any detector that achieves vanishing false-alarm probability.

### III. ASYNCHRONOUS COMMUNICATION

**A. Model**

We review here the asynchronous communication setup of \( \ref{fig:async} \). The mathematical differences between this model and the previously developed pure detection model are highlighted at the end of this section.

Consider discrete-time communication over a discrete memoryless channel characterized by its finite input and output alphabets
\[
\mathcal{X} \cup \{\ast\} \quad \text{and} \quad \mathcal{Y},
\]
respectively, and transition probability matrix
\[
Q(y|x),
\]
for all \( y \in \mathcal{Y} \) and \( x \in \mathcal{X} \cup \{\ast\} \). Here \( \ast \) denotes the “idle” symbol and \( \mathcal{X} \) represents the set of input symbols that can be used for codebook design. Without loss of generality, we assume that for all \( y \in \mathcal{Y} \) there is some \( x \in \mathcal{X} \cup \{\ast\} \) for which \( Q(y|x) > 0 \).

Given \( B \geq 0 \) information bits to be transmitted, a codebook \( \mathcal{C} \) consists of
\[
M = 2^B
\]
codewords of length \( n \geq 1 \) composed of symbols from \( \mathcal{X} \).

A randomly and uniformly chosen message \( m \) is available at the transmitter at a random time \( \nu \), independent of \( m \), and uniformly distributed over \( \{1, \ldots, A_B\} \), where the integer
\[
A_B = 2^{\beta B}
\]
characterizes the asynchronism level between the transmitter and the receiver, and where the constant
\[
\beta \geq 0
\]
denotes the timing uncertainty per information bit (see Fig.\( \ref{fig:async} \)). While \( \nu \) is unknown to the receiver, \( A_B \) is known by both the transmitter and the receiver.

We consider one-shot communication, i.e., only one message arrives over the period \( \{1, 2, \ldots, A_B\} \). If \( A_B = 1 \), the channel is said to be synchronous.

Given \( \nu \) and \( m \), the transmitter chooses a time \( \sigma(\nu, m) \) to start sending codeword \( c^\nu(m) \in \mathcal{C} \) assigned to message \( m \). Transmission cannot start before the message arrives or after the end of the uncertainty window, hence \( \sigma(\nu, m) \) must satisfy
\[
\nu \leq \sigma(\nu, m) \leq A_B \quad \text{almost surely.}
\]

In the rest of the paper, we suppress the arguments \( \nu \) and \( m \) of \( \sigma \) when these arguments are clear from context.

Before and after the codeword transmission, i.e., before time \( \sigma \) and after time \( \sigma + n - 1 \), the receiver observes “pure noise,” specifically, conditioned on the event \( \{\nu = t\} \), \( t \in \{1, \ldots, A_B\} \), and on the message to be conveyed \( m \), the receiver observes independent channel outputs
\[
Y_1, Y_2, \ldots, Y_{A_B+n-1}
\]
distributed as follows. For
\[
1 \leq i \leq \sigma(t, m) - 1
\]
or
\[
\sigma(t, m) + n \leq i \leq A_B + n - 1,
\]
the \( Y_i \)’s are “pure noise” symbols, i.e.,
\[
Y_i \sim Q(\cdot|\ast).
\]
For \( \sigma \leq i \leq \sigma + n - 1 \)
\[
Y_i \sim Q(\cdot|c_\nu(m))
\]
where \( c_i(m) \) denotes the \( i \)th symbol of the codeword \( c^\nu(m) \).

The receiver in a synchronous setting operates according to a decoding function only, which is a map from the channel outputs to the message set. In the present context of asynchronous communication with a sampling constraint, decoding involves a three level procedure:

- a sampling strategy (in the sense of Definition\( \ref{def:sampling} \) defined on the channel output process,
- a stopping (decoding) time defined on the sampled process,
where sages) decoding error probability of a code sampling strategy, and a decoder (decision time and decoding function). Whenever clear from context, we often refer to a

Assumption: we make the assumption that ⋆ means that the transmitter can stay idle at no cost only if ⋆ ∈ X and starts being sent at time ν, and decodes at time τ. The receiver samples at the (random) times S_1, S_2, . . . and decodes at time τ based on past samples.

• a decoding function defined on the stopped sampled process.

Once the sampling strategy is fixed, the receiver decodes by means of a sequential test (τ, φ_τ) where τ denotes a stopping time with respect to the sampled sequence Y_{S_1}, Y_{S_2}, . . .

and where φ_τ denotes the decoding function

\[ φ_τ : Y^{|Ω_1|} \rightarrow \{1, 2, \ldots, M\} \]

\[ O_τ ↦ φ_τ(O_τ) . \]

with

\[ O_t \equiv \{Y_{S_i} : S_i ≤ t\} \]

denoting the set of observations until time t.

A code (C, S, (τ, φ_τ)) is defined as a codebook, a receiver sampling strategy, and a decoder (decision time and decoding function). Whenever clear from context, we often refer to a code using the codebook symbol C only, leaving out an explicit reference to the sampling strategy and to the decoder.

Definition 6 (Error probability). The maximum (over messages) decoding error probability of a code C is defined as

\[ \max \limits_{m} P_m(E_m), \]  

(5)

where

\[ P_m(E) \equiv \frac{1}{A} \sum_{t=1}^{A} P_{m,t}(E_m), \]

and where the subscripts “m, t” denote conditioning on the event that message m arrives at time ν = t, and where E_m denotes the error event that the decoded message does not correspond to m, i.e.,

\[ E_m \equiv \{φ_τ(O_τ) \neq m\} . \]  

(6)

Definition 7 (Cost of a Code). The (maximum) cost of a code C with respect to a cost function k : X → [0, ∞) is defined as

\[ K(C) \equiv \max \limits_{m} \sum_{i=1}^{n} k(c_i(m)). \]

Assumption: we make the assumption that * has zero cost and that all other symbols have strictly positive costs. This means that the transmitter can stay idle at no cost only if * ∈ X. When * ∉ X then k(x) > 0 for any x ∈ X.

The other cases—investigated in [3] under full sampling—are either trivial (when X contains two or more zero cost symbols) or arguably unnatural (X contains a zero cost symbol that differs from * or when * ∈ X and all X contains only nonzero cost symbols) and are omitted in this paper.

Definition 8 (Decoding delay). Given ε > 0, the (maximum) delay of a code C, denoted by d(C, ε), is defined as the minimum integer l such that

\[ \min \limits_{m} P_m(τ − ν ≤ l − 1) ≥ 1 − ε . \]

Definition 9 (Sampling rate of a code). Given ε > 0, the sampling rate of a code C, denoted by ρ(C, ε), is defined as the minimum r ≥ 0 such that

\[ \min \limits_{m} P_m(B_r) ≥ 1 − ε . \]

where B_r is defined in (1).

We now define capacity per unit cost under the constraint that the receiver has access to a limited number of channel outputs:

Definition 10 (Asynchronous Capacity per Unit Cost under Sampling Constraint). Fix β ≥ 0 and fix a sequence of non-increasing values \{ρ_B\} with 0 ≤ ρ_B ≤ 1.

R is an achievable rate per unit cost at timing uncertainty per information bit β and sampling rates \{ρ_B\} if there exists a sequence of codes \{C_B\} such that for all B large enough

1) C_B operates at timing uncertainty per information bit β;
2) the maximum error probability is at most ε_B;
3) the rate per unit cost

\[ \frac{B}{K(C_B)} \]

is at least R − ε_B;
4) the sampling rate satisfies ρ(C_B, ε_B) ≤ ρ_B;
5) the delay satisfies

\[ \frac{1}{B} \log(d(C_B, ε_B)) ≤ ε_B \]

for some sequence of nonnegative numbers \{ε_B\} such that ε_B → 0.

Capacity per unit cost, denoted by C(β, \{ρ_B\}), is defined as the supremum of achievable rates per unit cost. Capacity per unit cost under full sampling, i.e., C(β, \{ρ_B = 1\}), is simply denoted by C(β).

Let us emphasize the mathematical differences between the model considered in this section and the pure detection problem developed in Section III. While the nominal distribution P_0 corresponds to the pure noise distribution Q_*, the posterior distribution here is no longer unique as it depends on the sent message. Furthermore, the goal of the receiver is to produce a reliable message estimate with a short detection delay. Because of this, the receiver, in addition to a stopping rule uses an isolation rule. Finally notice that here the instant of the change may be tied to the change distribution since the instant of the change σ = σ(ν, m) may be chosen as a function of the message to be conveyed.
Notational convention

Similarly as the convention for $d_n$ and $\rho_n$, we use $d_B$ and $\rho_B$ instead of $d(C_B, \varepsilon_B)$ and $\rho(C_B, \varepsilon_B)$, respectively, leaving out any explicit reference to $C_B$ and to the sequence of nonnegative numbers $\{\varepsilon_B\}$ which we assume satisfies $\varepsilon_B \rightarrow 0$.

The results in the next section adopt a less neutral and more “communication” type of notation and express key quantities such as entropy, mutual information, and divergence using the standard random variable convention. For instance $D(Y_1||Y_2)$ shall refer to the divergence between the distributions of random variables $Y_1$ and $Y_2$, respectively.

B. Results

Capacity per unit cost under full sampling is given by the following theorem:

**Theorem 3** (Capacity, full sampling, Theorem 1 [8]). For any $\beta \geq 0$

$$C(\beta) = \max_X \left\{ \frac{I(X;Y) + D(Y||Y_\beta)}{\mathbb{E}[k(X)]} : X \in \mathcal{P}(R) \right\}, \tag{7}$$

where $\max_X$ denotes maximization with respect to the channel input distribution $P_X$, where $(X,Y) \sim P_X(:,Q(\cdot))$, where $Y_\beta$ denotes the random output of the channel when the idle symbol $\ast$ is transmitted (i.e., $Y_\ast \sim Q(\cdot|\ast)$), where $I(X;Y)$ denotes the mutual information between $X$ and $Y$, and where $D(Y||Y_\beta)$ denotes the divergence between the distributions of $Y$ and $Y_\beta$.

Theorem 3 characterizes capacity per unit cost under full output sampling and over codes whose delay grow subexponentially with $B$. As it turns out, the same capacity per unit cost can be achieved with linear delay and sparse output sampling.

Define

$$n^*_B(\beta, R) \overset{def}{=} \frac{B}{R \mathbb{E} \left[ k(X) : X \in \mathcal{P}(R) \right]} = \Theta(B),$$

where

$$\mathcal{P}(R) \overset{def}{=} \left\{ X : \min \left\{ \frac{I(X;Y) + D(Y||Y_\beta)}{\mathbb{E}[k(X)]} : X \in \mathcal{P}(R) \right\} \geq R \right\}. \tag{9}$$

The quantity $n^*_B(\beta, R)$ plays the same role as $n^*(\alpha)$ in Section III and quantifies the minimum detection delay as a function of the asynchronism level and rate per unit cost, under full sampling:

**Theorem 4** (Capacity, minimum delay, constant sampling rate, Theorem 3 [11]). Fix $\beta \geq 0$ and $R \in (0, C(\beta))$. For any codes $\{C_B\}$ that achieve rate per unit cost $R$ at timing uncertainty $\beta$ and sampling rate $\rho_B = 1$ we have

$$\liminf_{B \rightarrow \infty} \frac{d_B}{n^*_B(\beta, R)} \geq 1.$$ 

Furthermore, for any $\rho \in (0, 1]$, there exist codes $\{C_B\}$ that achieve rate $R$ at timing uncertainty $\beta$ and sampling rate $\rho_B = \rho$ and such that

$$\limsup_{B \rightarrow \infty} \frac{d_B}{n^*_B(\beta, R)} \leq 1.$$ 

The first part of Theorem 4 says that under full sampling the minimum delay achieved by rate $R \in (0, C(\beta))$ codes is $n^*_B(\beta, R)$. The second part of the theorem says that this minimum delay can be achieved even if the receiver samples only a constant fraction $\rho > 0$ of the channel outputs.

The natural question is then “What is the minimum sampling rate of codes that achieve rate $R$ and minimum delay $n^*_B(\beta, R)$?” Our main result says that this minimum sampling rate essentially decreases as $1/B$:

**Theorem 5** (Capacity, minimum delay, minimum sampling rate). Consider a sequence of codes $\{C_B\}$ that operate under timing uncertainty per information bit $\beta > 0$. If

$$\rho_B d_B = o(1),$$

the receiver does not even sample a single component of the sent codeword with probability bounded away from zero. Hence, the average error probability is bounded away from zero whenever $d_B = O(B)$ and $\rho_B = o(1/B)$.

Moreover, for any $R \in (0, C(\beta))$, there exist codes $\{C_B\}$ that achieve rate $R$ at timing uncertainty $\beta$ and delay

$$d_B \leq n^*_B(\beta, R)(1 + o(1))$$

as long as the sampling rate satisfies $\rho_B = \omega(1/B)$.

If $R > 0$, the minimum delay is order $O(B)$ by Theorem 4 and (8) and to achieve this minimum delay it is thus necessary that $\rho_B = \Omega(1/B)$.

IV. Proofs

Typicality convention

A length $q \geq 1$ sequence $v^q$ over $\mathcal{V}^q$ is said to be typical with respect to some distribution $\bar{P}$ over $\mathcal{V}$ if

$$||\hat{P}_{v^q} - \bar{P}|| \leq q^{-1/3}$$

where $\hat{P}_{v^q}$ denotes the empirical distribution (or type) of $v^q$.

Typical sets have large probability in the sense that

$$P^q(||\hat{P}_{v^q} - \bar{P}|| \leq q^{-1/3}) \rightarrow 1 \quad (q \rightarrow \infty)$$

as can be verified from Chebyshev’s inequality—in the above expressions $P^q$ denotes the $q$-fold product distributions of $\bar{P}$.

Moreover, for any distribution $\bar{P}$ over $\mathcal{V}$ we have

$$P^q(||\hat{P}_{v^q} - \bar{P}|| \leq q^{-1/3}) \leq 2^{-q(D(\bar{P}||P) - o(1))}. \tag{12}$$

About rounding

Throughout computations we ignore issues related to the rounding of non-integer quantities as they play no role asymptotically.

A. Proof of Theorem 4

The proof of Theorem 4 is essentially a Corollary of Theorem 4. We sketch the main arguments.

\[ || \cdot || \] refers to the $L_1$-norm.
Fig. 2. Parsing of the entire sequence of size $A_n + n - 1$ into $r_n$ blocks of length $d_n$, one of which is generated by $P_1$, while all the others are generated by $P_3$.

1) : To establish achievability of $D(P_1 || P_0)$ one uses the same sequential typicality detection procedure as in the achievability of [4] Theorem. For the converse argument, we use essentially the same arguments as for the converse of [4] Theorem. For this latter setting, achieving $\alpha$ means that we can drive the probability of the event $\{\tau \neq \nu + n - 1\}$ to zero. Although this performance metric differs from ours—vanishing probability of false-alarm and subexponential delay—a closer look at the converse argument of [4] Theorem reveals that if $\alpha > D(P_1 || P_0)$ there are exponentially many sequences of length $n$ that are “typical” with respect to the posterior distribution. This, in turn, implies that either the probability of false-alarm is bounded away from zero or that the delay is exponential.

2) : Consider stopping times $\{\tau_n\}$ that achieve delay $\{d_n\}$, and vanishing false-alarm probability (recall the notational conventions for $d_n$ at the end of Section II-A).

We define the “effective process” $\{\tilde{Y}_i\}$ as the process whose change has duration $\min\{d_n, n\}$ (instead of $n$).

Effective output process: The effective process is defined as $\tilde{Y}_i = Y_i \ 1 \leq i \leq \nu + \min\{d_n, n\} - 1$

$\{\tilde{Y}_i : \nu + \min\{d_n, n\} \leq i \leq A_n + n - 1\}$ i.i.d. $P_0$ process

Hence, the effective process differs from the true process over the period $\{1, 2, \ldots, \tau\}$ only when $\{\tau \geq \nu + d_n\}$ with $d_n < n$.

Genie aided statistician: A genie aided statistician observes the entire effective process (of duration $A_n + n - 1$) and is informed that the change occurred over $r_n$ consecutive (disjoint) blocks of duration $d_n$ as shown in Fig. 2.

The genie aided statistician produces a time interval of size $d_n$ which corresponds to an estimate of the change in distribution and is declared to be correct only if this interval corresponds to the change in distribution.

Observe that since $\tau$ achieves false-alarm probability $\varepsilon_n$ and delay $d_n$ on the true process $\{Y_i\}$, the genie aided statistician achieves error probability at most $2\varepsilon_n$. The extra $\varepsilon_n$ comes from the fact $\tau$ stops after time $\nu + d_n - 1$ (on $\{Y_i\}$) with probability at most $\varepsilon_n$. Therefore, with probability at most $\varepsilon_n$ the genie aided statistician observes a process that may differ from the true process.

By using the same arguments as for the converse of [4] Theorem but on process $\{\tilde{Y}_i\}$ parsed into consecutive slots of size $d_n$, one shows that if

$$\liminf_{n \to \infty} \frac{d_n}{n^*(\alpha)} < 1$$

then the error probability of the genie aided decoder tends to one.

3) : To establish achievability apply the same sequential typicality test as in the achievability part of [4] Theorem. ■

B. Proof of Theorem 2: Converse

Consider a sequence of detectors $\{(S_n, \tau_n)\}$ that achieves, for some $\varepsilon_n \to 0$, sampling rate $\{\rho_n\}$, communication delay $d_n$, and error probability $\varepsilon_n$ (recall the notational conventions for $d_n$ and $\rho_n$ at the end of Section II-A).

We now show that if

$$\rho_n = o(1/n),$$

then the detectors do not see even one $P_1$-generated sample with probability asymptotically bounded away from zero. This, in turn, implies that the delay is exponential (whenever the false-alarm probability vanishes).

We lower bound the probability of sampling only from the nominal distribution $P_0$ as

$$\mathbb{P}\{(\nu, \nu + 1, \ldots, \nu + n - 1) \cap S_\tau = \emptyset\}$$

$$\geq \mathbb{P}\{(\nu, \nu + 1, \ldots, \nu + n - 1 \land k) \cap S_\tau = \emptyset\}$$

$$\geq \sum_{t} \sum_{s_t} \mathbb{P}(\tau = t, S_\tau = s_t, \nu = j)$$

$$\geq \sum_{t} \mathbb{P}(\tau = t) \sum_{s_t} \mathbb{P}(S_\tau = s_t) \sum_{j} \mathbb{P}(\nu = j)$$

$$\geq \mathbb{P}(\tau = t) \sum_{s_t: |s_t| \leq k/2n} \mathbb{P}(S_\tau = s_t) \frac{k - |s_t|n}{A_n}$$

$$\geq \frac{k}{2A_n} \mathbb{P}_0(|S_k| \leq k/2n)$$

where we defined the set of indices $\partial_{k, s_t}$ as

$$\partial_{k, s_t} = \{j : \{j, j + 1, \ldots, j + n - 1 \land k\} \cap s_t = \emptyset\}.$$ 

Equality (a) holds for any $k \geq 1$. For (b), $s_t \subseteq \{1, 2, \ldots, A_n + n - 1\}$ ranges over all sampling strategies that have non-zero probability when conditioned on $\tau = t$. Equality (c) holds by noting that event $\{\tau = t\}$ is a function of $\{Y_i, i \in s_t\}$ and that $\nu = j \notin s_t$ for all $j \in \partial_{k, s_t}$. Hence samples in $s_t$ are all distributed according to the nominal distribution $\mathbb{P}_0$ ($P_0$-product distribution).

By assumption on the sampling rate we have

$$1 - \varepsilon_n \leq \mathbb{P}(|S_\tau| \leq \tau \rho_n)$$

which implies

$$1 - \varepsilon_n \leq \mathbb{P}(|S_\tau| \leq (1 + o(1))A_n \rho_n)$$

$8a \land b = \min\{a, b\}$
for any \( \varepsilon > 0 \) and \( n \) large enough since \( \tau \leq A_n + n - 1 \). Now, for any fixed \( 1 \leq k \leq A_n \)

\[
\mathbb{P}(|S_r| \leq (1 + o(1))A_n \rho_n) \\
\leq \mathbb{P}(|S_r| \leq (1 + o(1))A_n \rho_n | \nu > k) \mathbb{P}(\nu > k) \\
+ \mathbb{P}(\nu \leq k) \\
\leq \mathbb{P}(|S_k| \leq (1 + o(1))A_n \rho_n | \nu > k) \mathbb{P}(\nu > k) \\
+ \mathbb{P}(\nu \leq k) \\
= \mathbb{P}_0(|S_k| \leq (1 + o(1))A_n \rho_n) (A_n - k) / A_n \\
+ k / A_n
\]

which is strictly positive. From (18) we get

\[
\mathbb{P}_0(|S_k| \leq (1 + o(1))A_n \rho_n) \geq \frac{1 - 2\varepsilon}{1 - \varepsilon}.
\]

From (14) and the definition of \( k_n \) we deduce that

\[
k_n / 2n \geq A_n \rho_n
\]

for \( n \) large enough. Therefore, from (19)

\[
\mathbb{P}_0(|S_k| \leq k_n / 2n) \geq \frac{1 - 2\varepsilon}{1 - \varepsilon}
\]

for \( n \) large enough. Hence, from (15)

\[
\mathbb{P}(\{\nu, \nu + 1, \ldots, \nu + n - 1\} \cap S_r = 0) \geq \frac{1 - 2\varepsilon}{1 - \varepsilon} \frac{\varepsilon}{2}
\]

which is strictly positive for any \( \varepsilon \in (0, 1/2) \). Therefore, with probability bounded away from zero the detector will sample only from the nominal distribution. This, as we now show, implies that the delay is exponential.

On the one hand, assuming a vanishing false-alarm probability we have for any constant \( \eta \in (0, 1) \)

\[
o(1) \geq \mathbb{P}(\tau < \nu) \\
\geq \mathbb{P}(\tau < \eta A_n | \nu \geq \eta A_n) (1 - \eta) \\
= \mathbb{P}_0(\tau < \eta A_n) (1 - \eta).
\]

This implies

\[
\mathbb{P}_0(\tau < \eta A_n) \leq o(1),
\]

and, therefore,

\[
\mathbb{P}(\tau \geq \eta A_n) \geq \mathbb{P}(\tau \geq \eta A_n | \nu \geq \eta A_n) (1 - \eta) \\
= \mathbb{P}_0(\tau \geq \eta A_n) (1 - \eta) \\
= 1 - \eta - o(1).
\]

Now, define events

\begin{align*}
&\mathcal{A}_1: \text{the detector stops at a time } \geq \eta A_n, \\
&\mathcal{A}_2: \{|S_r| \leq \tau \rho_n\}, \\
&\mathcal{A}_3: \text{all samples taken up to time } \tau \text{ are distributed according to } P_0,
\end{align*}

and let \( \mathcal{A} \triangleq \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3. \)

From (20), (21), and (16), for any \( \varepsilon \in (0, 1/2) \) one can pick \( \eta \in (0, 1) \) small enough such that

\[
\liminf_{n \to \infty} \mathbb{P}(\mathcal{A}) > 0.
\]

We now argue that when event \( \mathcal{A} \) happens, the detector misses the change which might have occurred, say, before time \( \eta A_n / 2 \), thereby implying a delay \( \Theta(A_n) \) since \( \tau \geq \eta A_n \) on \( \mathcal{A} \).

When event \( \mathcal{A} \) happens, the detector takes \( o(A_n/n) \) samples (this follows from event \( \mathcal{A}_2 \) since by assumption \( \rho_n = o(1/n) \)). Therefore, within \( \{1, 2, \ldots, \eta A_n/2\} \) there are at least \( \eta A_n/2 - o(A_n) \) time intervals of length \( n \) that are unsampled. Each of these corresponds to a possible change. Therefore, when event \( \mathcal{A} \) happens, with probability at least \( \eta/2 - o(1) \) the change happens before time \( \eta A_n/2 \) whereas \( \tau \geq \eta A_n \). This implies that the delay is exponential in \( n \) since the probability of \( \mathcal{A} \) is asymptotically bounded away from zero.

\section{C. Proof of Theorem 2 Achievability}

We describe a detection procedure that asymptotically achieves minimum delay \( n^*(\alpha) \) and any sampling rate that is \( \omega(1/n) \) whenever \( \alpha \in (0, D(P_0 || P_0)) \).

Fix \( \alpha \in (0, D(P_1 || P_0)) \) and pick \( \varepsilon > 0 \) small enough so that

\[
n^*(\alpha)(1 + 2\varepsilon) \leq n.
\]

Suppose we want to achieve some sampling rate \( \rho_n = f(n) / n \) where \( f(n) = \omega(1) \) is some arbitrary increasing function (upper bounded by \( n \) without loss of generality). Define

\[
\bar{\Delta}(n) \overset{\text{def}}{=} n / f(n)^{1/3}
\]

\( s \)-instants \( \{t = j \bar{\Delta}(n), j \in \mathbb{N}^*\} \),

and recursively define

\[
\Delta_0(n) \overset{\text{def}}{=} f(n)^{1/3} \\
\Delta_i(n) \overset{\text{def}}{=} \min \{2^{\Delta_{i-1}(n)}, n^*(\alpha)(1 + \varepsilon)\}
\]

for \( i = 1, 2, \ldots, \ell \) where \( \ell \) denotes the smallest integer \( i \) such that \( \Delta_i(n) = n^*(\alpha)(1 + \varepsilon) \). The constant \( c \) in the definition of \( \Delta_i(n) \) is any fixed value so that

\[
0 < c < D(P_1 || P_0).
\]

The detector starts sampling in phases at the first \( s \)-instant (i.e., at time \( t = \Delta(n) \)) as follows:

1. **Preamble detection (phase zero):** Take \( \Delta_0(n) \) consecutive samples and check if they are typical with respect to \( P_1 \). If the test turns negative, meaning that \( \Delta_0(n) \) samples are not typical, skip samples until the next \( s \)-instant and repeat the procedure i.e., sample and test \( \Delta_0(n) \) observations. If the test turns positive, move to confirmation phases.

2. **Preamble confirmations (variable duration, \( \ell - 1 \) phases at most):** Take another \( \Delta_1(n) \) consecutive samples and check if they are typical with respect to \( P_1 \). If the test turns negative skip samples until the next \( s \)-instant and repeat Phase zero (and test \( \Delta_0(n) \) samples). If the test turns positive, perform a second confirmation
phase with $\Delta_1(n)$ replaced with $\Delta_2(n)$, and so forth. (Each confirmation phase is performed on a new set of samples.) If $\ell - 1$ consecutive confirmation phases (with respect to the same $s$-instant) turn positive, the receiver moves to the full block sampling phase.

3 Full block sampling ($\ell$-th phase): Take another

$$\Delta_\ell(n) = n^* (\alpha) (1 + \varepsilon)$$

samples and check if they are typical with respect to $P_1$. If they are typical, stop, otherwise skip samples until the next $s$-instant and repeat Phase zero. If by time $A_n + n - 1$ no sequence is found to be typical, stop.

For the false-alarm we have

$$P(\tau < \nu) \leq 2^{\alpha n} \cdot 2^{-n^*(\alpha)(1+\varepsilon)} (D(P_1||P_0) - o(1))$$

$$= 2^{-n\alpha \Theta(\varepsilon)}$$

(23)

because whenever the detector stops, last

$$n^*(\alpha) (1 + \varepsilon)$$

samples are necessarily typical with respect to $P_1$. Therefore the inequality (23) follows from (12) and a union bound over time indices. The equality in (23) holds by definition of $n^*(\alpha)$ (see (3)).

For the delay we get

$$P(\tau \leq \nu + (1 + 2\varepsilon)n^*(\alpha)) = 1 - o(1).$$

(24)

To see this note that

$$\Delta(n) + \sum_{i=0}^{\ell} \Delta_i(n) \leq (1 + 2\varepsilon)n^*(\alpha)$$

for $n$ large enough and that, by (12), the series of $\ell + 1$ hypothesis test will turn positive with probability $1 - o(1)$ when samples are distributed according to $P_1$.

Since $\varepsilon$ can be made arbitrarily small, from (23) and (24) we deduce that the detector achieves minimum delay (see Theorem 1 Claim 2).

To show that the above detection procedure achieves sampling rate

$$\rho_n = f(n)/n$$

we need to establish that

$$P(\frac{|S_\tau|}{\tau} \geq \rho_n) \to 0.$$

(25)

To prove this we first compute the sampling rate of the detector when entirely run over an i.i.d.-$P_0$ sequence. As should be intuitively clear, this will essentially give us the desired result since the duration of the change $n$ is negligible with respect to $A_n$ and since $\nu$ is uniformly distributed over $\{1, 2, \ldots, A_n\}$.

We start by computing the expected number of samples $N$ taken by the detector at any given $s$-instant when the detector is started at that specific $s$-instant and when the observations are all i.i.d. $P_0$. Obviously, by stationarity this expectation does not depend on the $s$-instant. We have

$$E_0 N = \Delta_0(n) + \sum_{i=0}^{\ell-1} p_i \cdot \Delta_{i+1}(n)$$

(26)

where $p_i$ denotes the probability that the $i$-th confirmation phase turns positive given that the detector is in the $i$-th confirmation phase.

From 12

$$p_i \leq 2^{-\alpha_n (D(P_1||P_0) - o(1))}$$

hence, using the definition of $\Delta_1(n)$ we deduce that

$$E_0 N_s = \Delta_0(n) (1 + o(1)).$$

(27)

Therefore, the expected number of samples taken by the detector up to any given time $t$ under $P_0$ can be upper bounded as

$$E_0 |S_t| \leq \frac{t}{\Delta(n)} \Delta_0(n) (1 + o(1))$$

$$= \frac{t f(n)^{2/3}}{n} (1 + o(1)).$$

(28)

This, as we now show, implies that the detector has the desired sampling rate. We have

$$P(\frac{|S_\tau|}{\tau} \geq \rho_n)$$

$$\leq P(\frac{|S_\tau|}{\tau} \geq \rho_n, \nu \leq \tau \leq \nu + (1 + 2\varepsilon)n^*(\alpha))$$

$$+ 1 - P(\nu \leq \tau \leq \nu + (1 + 2\varepsilon)n^*(\alpha))$$

$$\leq P(\frac{|S_\tau|}{\tau} \geq \rho_n, \nu \leq \tau \leq \nu + 2n)$$

$$+ 1 - P(\nu \leq \tau \leq \nu + (1 + 2\varepsilon)n^*(\alpha))$$

(29)

where the second inequality holds for $\varepsilon$ small enough by the definition of $n^*(\alpha)$.

The fact that

$$1 - P(\nu \leq \tau < \nu + (1 + 2\varepsilon)n^*(\alpha)) = o(1)$$

(30)

follows from (28) and (24). For the first term on the right-hand side of the second inequality in (29) we have

$$P(\frac{|S_\tau|}{\tau} \geq \rho_n, \nu \leq \tau \leq \nu + 2n)$$

$$\leq P(\frac{|S_{\nu+2n}|}{\nu+2n} \geq \nu \rho_n)$$

$$\leq P(\frac{|S_{\nu-1}|}{\nu-1} \geq \nu \rho_n - 2n - 1).$$

(31)

Since $S_{\nu-1}$ represents sampling times before the change (the underlying process is thus i.i.d. $P_0$), assuming

$$\nu \geq \sqrt{A_n} = 2^{an}$$

(32)

9Boundary effects due to the fact that $A_n$ need not be a multiple of $\Delta_n$ play no role asymptotically and thus are ignored.
we have
\[
\mathbb{P}(|S_{\nu-1}| \geq \nu\rho_n - 2n - 1|\nu) \leq \frac{\mathbb{E}_0[S_{\nu}]}{\nu\rho_n - 2n - 1}
\leq \frac{f(n)2^{1/3}(1 + o(1))}{n(\rho_n - (2n + 1)/\nu)}
\leq \frac{f(n)2^{1/3}(1 + o(1))}{\nu\rho_n(1 - o(1))}
= \frac{(1 + o(1))}{f(n)^{1/3}(1 - o(1))}
= o(1)
\]
where the second inequality holds by (28); where the third inequality holds by (32) and because \(\nu \geq \sqrt{A_n}\) and where the last two equalities hold by the definitions of \(\rho_n\) and \(f(n)\).

Removing the conditioning on \(\nu\),
\[
\mathbb{P}(|S_{\nu-1}| \geq \nu\rho_n - 2n - 1) \leq \mathbb{P}(\nu \geq \sqrt{A_n}) + \mathbb{P}(\nu < \sqrt{A_n}) = o(1)
\]
by (33) and since \(\nu\) is uniformly distributed over \(\{1, 2, \ldots, A_n\}\). Hence, from (31), the first term on the right-hand side of the second inequality in (29) vanishes.

This yields (25).

\[\text{Discussion}\]

There is obviously a lot of flexibility around the quickest detection procedure described in Section 11I-C. Its main feature is the multiple binary hypothesis test which rejects the hypothesis that a change occurred under pure noise as soon as possible while controlling false-alarms.

It may be tempting to simplify the detection procedure by considering, say, only two phases, the preamble phase and the full block phase. Such a scheme, which is similar in spirit to the one proposed in [1], would not work as it would produce a much higher level of false-alarm. We provide an intuitive justification for this thereby highlighting the role of the multiphase procedure.

Consider a two phase procedure, a preamble phase followed by a full block phase. Each time we switch to the second phase we take \(\Theta(n)\) samples. Therefore, if we want to achieve a vanishing sampling rate, then necessarily the probability of changing mode under pure noise should be \(o(1/n)\). By Sanov’s theorem, such a probability can be achieved only if the decision to change mode is based on \(\omega(\log n)\) samples taken over time windows of size \(\Theta(n)\). This translates into a sampling rate of \(\omega(\log n)/n\) at best, and we know that this is suboptimal since any sampling rate \(\omega(1/n)\) is achievable.

The reason a two-phase scheme does not yield a sampling rate lower than \(\omega(\log n)/n\) is that it is not progressive enough. To guarantee a vanishing sampling rate, the decision to switch to the full block phase should be based on at least \(\log(n)\) samples, which in turn yields a suboptimal sampling rate.

The important observation here is that the (average) sampling rate of the two-phase procedure essentially corresponds to the sampling rate of the first phase. And it is in this regime that it is decided when to switch to the full block phase and sample continuously for a long period of order \(n\). In the multiphase procedure, however, this is no longer the case. The sampling rate of the first phase does no more correspond to the sampling rate of the phase that can trigger the densest sampling mode.

Note also that the decision to switch to a higher sampling rate can happen more frequently than with a two-phase scheme because multiple phases give us the important ability to take far fewer than \(n\) samples when we make the decision to switch to a higher sampling rate. If the intermediate phase lengths are chosen appropriately, then we enter high sampling rates with low probability, and this does not affect the average sampling rate.

In general, the lengths and probability thresholds need to be chosen so that the sampling rate is dominated by the first phase. This translates into condition (26 and 27):
\[
\sum_{i=0}^{\ell-1} p_i \cdot \Delta_{i+1}(n) = o(\Delta_0(n)).
\]

\[\text{D. Converse of Theorem 5}\]

By using the same arguments as for the converse of Theorem 2 (until equation (20)) with \(n\) replaced by \(d_B\) one shows that if
\[
\rho_B d_B = o(1)
\]
the decoder does not even sample one component of the sent codeword with probability asymptotically bounded away from zero.

\[\text{E. Achievability of Theorem 5}\]

Fix \(\beta > 0\). We show that any \(\overline{R} \in (0, C(\beta)]\) is achievable with codes \(\{\mathcal{E}_B\}\) whose delays satisfy \(d(\mathcal{E}_B, \varepsilon_B) \leq n_B^*(\beta, \overline{R})(1 + o(1))\) whenever the sampling rate \(\rho_B\) is such that
\[
\rho_B = \frac{f(B)}{B}
\]
for some \(f(B) = \omega(1)\).

Let \(X \sim P\) be some channel input and let \(Y\) denote the corresponding output, i.e., \((X,Y) \sim P(\cdot)Q(\cdot|\cdot)\). For the moment we only assume that \(X\) is such that \(I(X;Y) > 0\). Further, we suppose that the codeword length \(n\) is linearly related to \(B\), i.e.,
\[
\frac{B}{n} = q
\]
for some fixed constant \(q > 0\). We shall specify this linear dependency later to accommodate the desired rate \(\overline{R}\). Further, let
\[
\tilde{f}(n) \overset{\text{def}}{=} f(q \cdot n)/q.
\]
and
\[ \hat{\rho}_n \overset{\text{def}}{=} \frac{\hat{f}(n)}{n}. \]

Hence, by definition we have
\[ \hat{\rho}_n = \rho_B. \]

Let \( a \) be some arbitrary fixed input symbol such that
\[ Q(\{a\}) \neq Q(\{\star\}). \]

Below we introduce the quantities \( \bar{\Delta}(n) \) and \( \bar{\Delta}_i(n), 1 \leq i \leq \ell, \) which are defined as in Section IV.C but with \( P_0 \) replaced with \( Q(\{\star\}), P_1 \) replaced with \( Q(\{a\}), f(n) \) replaced with \( \hat{f}(n) \), and \( n^*(\alpha) \) replaced with \( n. \)

1) Codewords: preamble followed by constant composition information symbols: Each codeword \( c^n(m) \) starts with a common preamble that consists of \( n-\bar{\Delta}(n) \) repetitions of symbol \( a. \) The remaining \( n-\bar{\Delta}(n) \) components
\[ c^n_{\bar{\Delta}(n)+1}(m) \]

of \( c^n(m) \) of each message \( m \) carry information and are generated as follows. For message 1, randomly generate length \( n-\bar{\Delta}(n) \) sequences \( x^{n-\bar{\Delta}}(n) \) i.i.d. according to \( P \) until when \( x^{n-\bar{\Delta}}(n) \) is typical with respect to \( P \). In this case we let
\[ c^n_{\bar{\Delta}(n)+1}(1) \overset{\text{def}}{=} x^{n-\bar{\Delta}(n)}, \]
move to message 2, and repeat the procedure until when a codeword has been assigned to each message.

From (11), for any fixed \( m \) no repetition will be required to generate \( c^n_{\bar{\Delta}(n)+1}(m) \) with probability tending to one as \( n \to \infty. \) Moreover, by construction codewords are essentially of constant composition—i.e., each symbol appears roughly the same number of times—and have cost
\[ nE[k(X)](1+o(1)) \]
as \( n \to \infty. \)

2) Codeword transmission time: Define the set of start instants
\[ s\text{-instants} \overset{\text{def}}{=} \{t = j\bar{\Delta}(n), j \in \mathbb{N}^*\}. \]

Codeword transmission start time \( \sigma(m, \nu) \) corresponds to the first \( s\)-instant \( \geq \nu \) (regardless of \( m \)).

3) Sampling and decoding procedures: The decoder first tries to detect the preamble by using a similar detection procedure as in the achievability of Theorem 2, then applies the standard message decoding isolation map.

Starting at the first \( s\)-instant (i.e., at time \( t = \bar{\Delta}(n) \)), the decoder samples in phases as follows.

1) Preamble test (phase zero): Take \( \Delta_0(n) \) consecutive samples and check if they are typical with respect to \( Q(\{a\}). \) If the test turns negative, the decoder skips samples until the next \( s\)-instant when it repeats the procedure. If the test turns positive, the decoder moves to the confirmation phases.

2) Preamble confirmations (variable duration, \( \ell - 1 \) phases at most): The decoder takes another \( \Delta_1(n) \) consecutive samples and checks if they are typical with respect to \( Q(\{a\}). \) If the test turns negative the decoder skips samples until the next \( s\)-instant when it repeats Phase zero (and tests \( \Delta_0(n) \) samples). If the test turns positive, the decoder performs a second confirmation phase based on new \( \Delta_2(n) \) samples, and so forth. If \( \ell - 1 \) consecutive confirmation phases (with respect to the same \( s\)-instant) turn positive, the decoder moves to the message sampling phase.

3) Message sampling and isolation (\( \ell \)-th phase): Take another \( n \) samples and check if among these samples there are \( n-\bar{\Delta}(n) \) consecutive samples that are jointly typical with the \( n-\bar{\Delta}(n) \) information symbols of one of the codewords. If one codeword is typical, stop and declare the corresponding message. If more than one codeword is typical declare one message at random. If no codeword is typical, the decoder stops sampling until the next \( s\)-instant and repeats Phase zero. If by time \( A_B + n - 1 \) no codeword is found to be typical, the decoder declares a random message.

4) Error probability: Error probability and delay are evaluated in the limit \( B \to \infty \) with \( A_B = 2\beta B \) and with
\[ q = \frac{B}{n} < \min \left\{ I(X;Y), \frac{I(X;Y) + D(Y\|Y_s)}{1+\beta} \right\}. \quad (36) \]

We first compute the error probability averaged over codebooks and messages. Suppose message \( m \) is transmitted. The decoding error event \( \mathcal{E}_m \) (see (6)) can be decomposed as
\[ \mathcal{E}_m \subseteq \mathcal{E}_{0,m} \cup m' \neq m (\mathcal{E}_{1,m'} \cup \mathcal{E}_{2,m'}), \quad (37) \]

where events \( \mathcal{E}_{0,m}, \mathcal{E}_{1,m'}, \) and \( \mathcal{E}_{2,m'} \) are defined as

- \( \mathcal{E}_{0,m} \): at the \( s\)-instant corresponding to \( \sigma \), the preamble test phase or one of the preamble confirmation phases turns negative, or \( c^n_{\bar{\Delta}(n)+1}(m) \) is not found to be typical by time \( \sigma + n - 1; \)
- \( \mathcal{E}_{1,m'} \): the decoder stops at a time \( t < \sigma \) and declares \( m'; \)
- \( \mathcal{E}_{2,m'} \): the decoder stops at a time \( t \) between \( \sigma \) and \( \sigma + n - 1 \) (including \( \sigma \) and \( \sigma + n - 1 \)) and declares \( m'. \)

From Sanov’s theorem,
\[ \mathbb{P}_m(\mathcal{E}_{0,m}) = \varepsilon_1(B) \quad (38) \]

where \( \varepsilon_1(B) = o(1). \) Note that this equality holds pointwise (and not only on average over codebooks) for any specific (non-random) codeword \( c^n(m) \) since, by construction, they all satisfy the constant composition property
\[ ||\hat{P}_{\bar{\Delta}+1}(m) - P|| \leq (n-\bar{\Delta})^{-1/3} = o(1) \quad (39) \]
as \( n \to \infty. \)

Using analogous arguments as in the achievability of Proof of Theorem 1, we obtain the upper bounds
\[ \mathbb{P}_m(\mathcal{E}_{1,m'}) \leq 2^{2B} \cdot 2^{-n(I(X;Y)+D(Y\|Y_s)-o(1))} \]
and
\[ \mathbb{P}_m(\mathcal{E}_{2,m'}) \leq 2^{-n(I(X;Y)-o(1))} \]
which are both valid for any fixed \( \varepsilon > 0 \) provided that \( B \) is large enough. Hence from the union bound

\[
\mathbb{P}_m(\mathcal{E}_{1,m'} \cup \mathcal{E}_{2,m'}) \leq 2^{-n(I(X;Y) - \varepsilon(1))} + 2^B \cdot 2^{-n(I(X;Y) + D(Y||Y_\ast) - \varepsilon(1))}.
\]

Taking a second union bound over all possible wrong messages, we get

\[
\mathbb{P}_m(\cup_{m' \neq m}(\mathcal{E}_{1,m'} \cup \mathcal{E}_{2,m'})) \leq 2^B \cdot 2^{-n(I(X;Y) - \varepsilon(1))} + 2^B \cdot 2^{-n(I(X;Y) + D(Y||Y_\ast) - \varepsilon(1))} \overset{\text{def}}{=} \varepsilon_2(B)
\]

(40)

where \( \varepsilon_2(B) = o(1) \) because of (36).

Combining (37), (38), (40), we get from the union bound

\[
\mathbb{P}_m(\mathcal{E}_m) \leq \varepsilon_1(B) + \varepsilon_2(B) = o(1)
\]

(41)

for any \( m \).

We now show that the delay of our coding scheme is at most \( n(1 + o(1)) \). Suppose codeword \( c^n(m) \) is sent. If \( \tau > \sigma + n \)

then necessarily \( c^n_{\Delta+1}(m) \) is not typical with the corresponding channel outputs. Hence

\[
\mathbb{P}_m(\tau - \sigma \leq n) \geq 1 - \mathbb{P}_m(\mathcal{E}_{0,m}) = 1 - \varepsilon_1(B)
\]

by (38). Since \( \sigma \leq \nu + \Delta(n) \) and \( \Delta(n) = o(n) \) we get

\[
\mathbb{P}_m(\tau - \nu \leq n(1 + o(1))) \geq 1 - \varepsilon_1(B).
\]

Since this inequality holds for any codeword \( c^n(m) \) that satisfies (39), the delay is no more than \( n(1 + o(1)) \) (see Definition 8). Furthermore, from (41) there exists a specific non-random code \( c \) whose error probability, averaged over messages, is less than \( \varepsilon_1(n) + \varepsilon_2(n) = o(1) \) whenever condition (35) is satisfied. Removing the half of the codewords with the highest error probability, we end up with a set \( \mathcal{E}' \) of \( 2^{B-1} \) codewords whose maximum error probability satisfies

\[
\max_m \mathbb{P}_m(\mathcal{E}_m) \leq o(1)
\]

(43)

whenever condition (36) is satisfied.

Since any codeword has cost \( n\mathbb{E}[k(X)](1 + o(1)) \), condition (36) is equivalent to

\[
R < \min \left\{ \frac{I(X;Y)}{\mathbb{E}[k(X)](1 + o(1))}, \frac{I(X;Y) + D(Y||Y_\ast)}{\mathbb{E}[k(X)](1 + o(1))(1 + \beta)} \right\}
\]

(44)

where

\[
R \overset{\text{def}}{=} \frac{B}{K(c')}
\]

denotes the rate per unit cost of \( c' \).

Thus, to achieve a given \( R \in (0, C(\beta)) \) it suffices to choose the input distribution and the codeword length as

\[
X = \arg \max \{ \mathbb{E}[k(X')] : X' \in \mathcal{P}(R) \}
\]

(11)

Recall that \( B/n \) is kept fixed and \( B \to \infty \).

\[\text{(see (8) and (9)). By a previous argument the corresponding delay is no larger than } n_B^2(\beta, R) (1 + o(1)).\]

5) Sampling rate: For the sampling rate, a very similar analysis as in the achievability of Theorem 2 (see from equation (23) onwards with \( f(n), \rho_n, n'/(\beta, R) \), and \( A_k \) replaced with \( f(n), \rho_n, n'/(\beta, R) \), and \( A_k \), respectively) shows that

\[
\mathbb{P}_m(\mathcal{S}_1/\tau \geq \rho_B) = \mathbb{P}_m(\mathcal{S}_1/\tau \geq \rho_n) \overset{B,n \to \infty}{\longrightarrow} 0.
\]

(45)

Note that the arguments that establish (45) rely only on the preamble detection procedure. In particular, they do not use (44) and hold for any codeword length \( n_B \) as long as \( n_B = \Theta(B) \).

\section*{References}

[1] A. Tchamkerten, V. Chandar, and G. Caire, “Energy and sampling constrained asynchronous communication,” Information Theory, IEEE Transactions on, vol. 60, no. 12, pp. 7686–7697, Dec 2014.

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, 2nd edition. MIT Press, McGraw-Hill Book Company, 2000.

[3] A. Tchamkerten, A. Khisti, and G. Wornell, “Information theoretic perspectives on synchronization,” in Information Theory, 2006 IEEE International Symposium on. IEEE, 2006, pp. 371–375.

[4] V. Chandar, A. Tchamkerten, and G. Wornell, “Optimal sequential frame synchronization,” Information Theory, IEEE Transactions on, vol. 54, no. 8, pp. 3725–3728, 2008.

[5] A. Tchamkerten, V. Chandar, and G. Wornell, “Communication under strong asynchronism,” Information Theory, IEEE Transactions on, vol. 55, no. 10, pp. 4508–4528, 2009.

[6] D. Wang, “Distinguishing codes from noise: fundamental limits and applications to sparse communication,” Master’s thesis, Massachusetts Institute of Technology, 2010.

[7] A. Tchamkerten, V. Chandar, and G. W. Wornell, “Asynchronous communication: Capacity bounds and suboptimality of training,” Information Theory, IEEE Transactions on, vol. 59, no. 3, pp. 1227–1255, march 2013.

[8] V. Chandar, A. Tchamkerten, and D. Tse, “Asynchronous capacity per unit cost,” Information Theory, IEEE Transactions on, vol. 59, no. 3, pp. 1213–1226, march 2013.

[9] Y. Polyanskiy, “Asynchronous communication: Exact synchronization, universality, and dispersion,” Information Theory, IEEE Transactions on, vol. 59, no. 3, pp. 1256–1270, 2013.

[10] N. Weinberger and N. Merhav, “Codeword or noise? exact random coding exponents for joint detection and decoding,” Information Theory, IEEE Transactions on, vol. 60, no. 9, pp. 5077–5094, Sept 2014.

[11] G. Lorden, “Procedures for reacting to a change in distribution,” The Annals of Mathematical Statistics, pp. 1897–1908, 1971.

[12] I. V. Nikiforov, “A generalized change detection problem,” IEEE Trans. Inform. Th., vol. 41, pp. 171–187, January 1995.

[13] B. Broder and S. Schwartz, “Quickest detection procedures and transient signal detection,” DTIC Document, Tech. Rep., 1990.

[14] C. Han, P. K. Willett, and D. A. Abraham, “Some methods to evaluate the performance of page’s test as used to detect transient signals,” Signal Processing, IEEE Transactions on, vol. 47, no. 8, pp. 2112–2127, 1999.

[15] T. L. Lai, “Information bounds and quick detection of parameter changes in stochastic systems,” Information Theory, IEEE Transactions on, vol. 41, pp. 171–187, January 1995.

[16] X. Fan, Y. Polyanskiy, and G. Wornell, “Bayesian quickest detection of parameter changes in stochastic systems under a sampling constraint,” submitted for publication.