Asynchronous Capacity per Unit Cost

Venkat Chandar, Aslan Tchamkerten, and David Tse

***Abstract*—The capacity per unit cost, or, equivalently, the minimum cost to transmit one bit, is a well-studied quantity under the assumption of full synchrony between the transmitter and the receiver. In many applications, such as sensor networks, transmissions are very bursty, with amounts of bits arriving infrequently at random times. In such scenarios, the cost of acquiring synchronization is significant and one is interested in the fundamental limits on communication without assuming *a priori* synchronization. In this paper, the minimum cost to transmit B bits of information asynchronously is shown to be equal to $(B + \bar{H})k_{\text{sync}}$, where k_{sync} is the synchronous minimum cost per bit, and where $\bar{H}$ is a measure of timing uncertainty equal to the entropy for most reasonable arrival time distributions. This result holds when the transmitter can stay idle at no cost and is a particular case of a general result which holds for arbitrary cost functions.**

***Index Terms*—asynchronous communication; bursty communication; capacity; capacity per unit cost; energy; error exponents; large deviations; sequential decoding; sparse communication; synchronization**

I. Introduction

Synchronization is an important component of any communication system. To understand the cost of synchronization, it is helpful to divide applications into two rough types. In the first type, transmission of data happens on a continuous basis. Examples are voice and video. The cost of initially acquiring synchronization, say by sending a pilot sequence, is relatively small in such applications because the cost is amortized over the many symbols transmitted. In the second type, transmissions are very bursty, with amounts of data transmitted once in a long while. Examples are sensor networks with sensor nodes transmitting measured data once in a while. The cost of acquiring synchronization is relatively more significant in such applications because the number of bits transmitted per burst is relatively small.

What is the fundamental limitation due to the lack of *a priori* synchrony between the transmitter and the receiver in bursty communication? While there has been a lot of research on specific synchronization algorithms, this question has only recently been pursued [1], [7], [6]. In their model, transmission of a message starts at a random time unknown to the receiver. The performance measure is the data rate: the number of bits in the message divided by the elapsed time between the instant information starts being sent and the instant it is decoded.

The data rate is a sensible performance metric for bursty communication if the information to be communicated is delay-sensitive. Then, maximizing the data rate is equivalent to minimizing the time to transmit the burst of data. In certain applications, however, the allowable delay may not be so tightly constrained, so the data rate is less relevant a measure than the *energy* needed to transmit the information. In this case, the minimum energy needed to transmit one bit of information is an appropriate fundamental measure. Thus, we are led to ask the following question: what is the impact of asynchrony on the minimum energy needed to transmit one bit of information?

This type of question falls into the general framework of *capacity per unit cost* [5], [8], where one is interested in characterizing the maximum number of bits that can be reliably communicated per unit cost of using the channel. Consider the following modification of the formulation in [7], [6] to study asynchronous capacity per unit cost.

There are B bits of information which needs to be communicated. The number B can be viewed as the size of a burst in the above scenario, with consecutive bursts occurring so infrequently that we can consider each burst in complete isolation. The B bits are coded and transmitted over a memoryless channel using a sequence of symbols that have costs associated with them. The rate $\boldsymbol{R}$ per unit cost is the total number of bits divided by the cost of the transmitted sequence.

The data burst arrives at a *random symbol time* ν, not known *a priori* to the receiver. Without knowing ν, the goal of the receiver is to reliably decode the information bits by observing the outputs of the channel. Although the receiver does not know ν, we assume that both the transmitter and the receiver know that ν lies in the range from 1 to A. The integer A characterizes the asynchronism level or the timing uncertainty between the transmitter and the receiver. At all times before and after the actual transmission, the receiver observes pure noise. The noise distribution corresponds to a special "idle symbol" $\star$ being sent across the channel.

The main result in this paper is a single-letter characterization of the asynchronous capacity per unit cost, or, equivalently, the minimum cost to transmit one bit of information. Under the further assumption that the idle symbol $\star$ is allowed to be used in the codewords and has zero cost, the result simplifies and admits a very simple interpretation: the minimum cost to transmit B bits of information asynchronously is

$$(B + \log A)k_{\text{sync}}, \tag{1}$$

where k_{sync} is the minimum cost to transmit one bit of

This work was supported in part by an Excellence Chair Grant from the French National Research Agency (ACE project). This work was presented in part at the IEEE International Symposium on Information Theory, Austin (Tx), USA, June 2010.

V. Chandar is with MIT Lincoln Laboratory, Lexington, MA 02420, USA. Email: vchandar@mit.edu.

A. Tchamkerten is with the Department of Communications and Electronics, Telecom ParisTech, 75634 Paris Cedex 13, France. Email: aslan.tchamkerten@telecom-paristech.fr.

D. Tse is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley CA 94729-1770, USA. Email: dtse@eecs.berkeley.edu.

information in the synchronous setting.[1] Thus, the timing uncertainty imposes an additional cost of $k_{\text{sync}} \log A$ as compared to the synchronous setting. Note that this result implies that the additional cost is significant only when the parameter $\log A$ is at least comparable to B.

Even though we do not have a *stringent* requirement on the delay from the time of data arrival to the time of decoding, a meaningful result cannot be obtained if there is *no* constraint at all. This can be seen by noting that the transmitter could always wait until the end of the arrival time interval (at time A) to transmit information. Then, there would no price to pay for the timing uncertainty since communication would *de facto* be synchronous. However, the delay incurred would be very large if A is very large. To avoid this undesirable situation, we impose the constraint that the delay should be *linear* in B. A delay linear in B is a natural constraint since it is of the same order as the delay incurred in the synchronous setting [8]. The expression (1) is the minimum cost achievable by any scheme subject to this delay constraint. Given this constraint, the start time of information transmission is highly random to the receiver and the additional cost is the cost needed to construct codewords that allow a decoder to resolve this uncertainty.

What happens when longer delays are allowed? First, we show that performance cannot be improved beyond (1) within the broad class of coding schemes whose delays are *sub-exponential* in B. Second, we show that when the allowable delay d scales exponentially with B (but is no larger than A, for otherwise the situation reduces to the synchronous setting mentioned above), the minimum cost to transmit B bits can be further reduced to

$$\left(B + \log \frac{A}{d}\right) k_{\text{sync}}.$$

Thus, in this more general case, the impact of asynchronism is significant when $\log(A/d)$ is at least of the order of B.

The above results are all proved under a uniform distribution on the arrival time ν. They can be generalized to a broad class of other distributions, with $\log A$ replaced by a quantity $\bar{H}$, which equals the entropy for most reasonable distributions.

It is worth mentioning that the asynchronism studied in this paper is due entirely to the random arrival time of the data and the desire to deliver that data within a certain delay constraint. One can think of this as *source* asynchronism. There is another type of asynchronism due to the lack of a common clock between the transmitter and the receiver. One can think of this as an example of *channel* asynchronism. We do not consider this type of asynchronism here. Hence, throughout the paper, we will assume both the transmitter and the receiver have access to a common clock. An interesting future direction would be to study the combined effect of source and channel asynchronism.

II. Model and Performance Criterion

Our model captures the following features:

- Information is available at the transmitter at a random time;
- The transmitter chooses when to start sending information;
- Outside the information transmission period, the transmitter stays idle and the receiver observes noise;
- The receiver decodes without knowing the information arrival time at the transmitter.

Communication is discrete-time, and carried over a discrete memoryless channel characterized by its finite input and output alphabets

$$\mathcal{X} \cup \{\star\} \quad \text{and} \quad \mathcal{Y},$$

respectively, and transition probability matrix

$$Q(y|x) \qquad x \in \mathcal{X} \cup \{\star\}, y \in \mathcal{Y}.$$

Here $\star$ denotes the special idle symbol, and $\mathcal{X}$ denotes the alphabet containing the symbols that can be used in the actual transmission of the data. $\mathcal{X}$ may or may not contain $\star$. We assume that no two different input symbols x and x' belonging to $\mathcal{X}$ have identical conditional distributions $Q(\cdot|x)$ and $Q(\cdot|x')$.[2]

Given B information bits to be transmitted, a codebook $\mathcal{C}$ consists of 2^B codewords of length n composed of symbols from $\mathcal{X}$. The message m arrives at the transmitter at a random time ν, independent of m, and uniformly distributed over $\{1, 2, \ldots, A\}$, where the integer $A \geq 1$ characterizes the *asynchronism level* between the transmitter and the receiver. Only one message arrives over the period $[1, 2, \ldots, A+n-1]$. If $A = 1$, the channel is said to be synchronous.

The transmitter chooses a time $\sigma(\nu, m)$ so that

$$\nu \leq \sigma(\nu, m) \leq A \qquad \text{almost surely}$$

to begin transmitting the codeword $c^n(m) \in \mathcal{C}$ assigned to message m. This means that the transmitter cannot start transmitting before the message arrives or after the end of the uncertainty window. It turns out that the possibility to choose σ as a function of both ν and m directly influences the cost to deliver this information by allowing to convey information through timing. In the rest of the paper, we suppress the arguments ν and m of σ when these arguments are clear from context.

Before and after codeword transmission, *i.e.*, before time σ and after time $\sigma + n - 1$, the receiver observes "pure noise." Specifically, conditioned on the event $\{\nu = t\}$, $t \in \{1, 2, \ldots, A\}$, and on the message to be conveyed m, the receiver observes independent symbols

$$Y_1, Y_2, \ldots, Y_{A+n-1}$$

distributed as follows. For

$$1 \leq i \leq \sigma(t, m) - 1$$

or

$$\sigma(t, m) + n \leq i \leq A + n - 1\,,$$

[1] In this paper, all logarithms are taken to base 2.

[2] This is without loss of generality, as two such symbols are identical for communication purposes, so we can consider the equivalent channel with one of these two symbols deleted from the symbol alphabet.

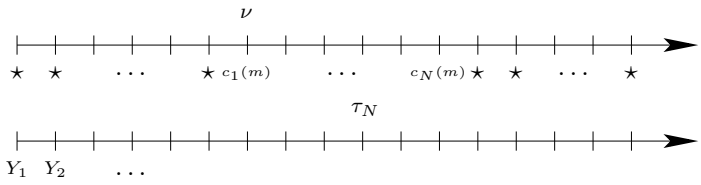

Fig. 1. Time representation of what is sent (upper arrow) and what is received (lower arrow). The "$\star$" represents the "idle" symbol. Message m arrives at time ν, starts being sent at time σ, and decoding occurs at time τ.

the Y_i's are distributed according to $Q(\cdot|\star)$. At any time $i \in \{\sigma, \sigma+1, \ldots, \sigma+n-1\}$, the distribution is

$$Q(\cdot|c_{i-\sigma+1}(m)),$$

where $c_i(m)$ denotes the ith symbol of the codeword $c^n(m)$.

Knowing the asynchronism level A, but not the value of ν, the receiver decodes by means of a sequential test (τ, ϕ), where τ is a stopping time, bounded by $A+n-1$, with respect to the output sequence $Y_1, Y_2, \ldots$ indicating when decoding happens, and where ϕ denotes a decision rule that declares the decoded message (see Fig. 1). Recall that a (deterministic or randomized) stopping time τ 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 $t \geq 1$. Given $\{\tau = t\}$, $t \in \{1, 2, \ldots, A+n-1\}$, the function ϕ outputs a message based on the past observations from time 1 up to time t.[3]

A "code" refers to a codebook $\mathcal{C}$ together with a decoder, *i.e.*, a sequential test (τ, ϕ). Throughout the paper, whenever clear from context, we often refer to a code using the codebook symbol $\mathcal{C}$ only, leaving out an explicit reference to the decoder.

The maximum (over messages) decoding error probability for a given code $\mathcal{C}$ is defined as

$$\mathbb{P}(\mathcal{E}|\mathcal{C}) \triangleq \max_m \frac{1}{A} \sum_{t=1}^{A} \mathbb{P}_{m,t}(\mathcal{E}|\mathcal{C}), \tag{2}$$

where the subscripts "m, t" indicate conditioning on the event that message m arrives at time $\nu = t$, and where $\mathcal{E}$ indicates the event that the decoded message does not correspond to the sent codeword, *i.e.*,

$$\mathcal{E} \triangleq \{\phi(Y^\tau) \neq M\}$$

where M denotes the random message to be transmitted.

Definition 1 (Cost Function)**.** *A cost function $k : \mathcal{X} \to [0, \infty)$ assigns a non-negative value to each channel input.*[4]

Definition 2 (Cost of a Code)**.** *The (maximum) cost of a code $\mathcal{C}$ is defined as*

$$K(\mathcal{C}) \triangleq \max_m \sum_{i=1}^{n} k(c_i(m)).$$

[3]To be more precise, ϕ is any $\mathcal{F}_\tau$-measurable function that takes values in the message set, where $\mathcal{F}_t$ is the sigma field generated by $Y_1, Y_2, \ldots, Y_t$.

[4]"Kost" is cost in German.

Definition 3 (Delay of a Code)**.** *Given $\varepsilon > 0$, the (maximum) delay of a code $\mathcal{C}$, denoted by $d(\mathcal{C}, \varepsilon)$, is defined as the smallest d such that*

$$\min_m \mathbb{P}_m(\tau - \nu \leq d - 1) \geq 1 - \varepsilon,$$

where $\mathbb{P}_m$ denotes the output distribution conditioned on the sending of message m.[5]

Throughout the paper, we often consider delays in the regime $\varepsilon \to 0$. In this case, we omit an explicit reference to ε. For instance, if $\{\mathcal{C}_B\}$ is such that $d(\mathcal{C}_B, \varepsilon_B) = O(B)$ for some $\{\varepsilon_B\}$ such that $\varepsilon_B \to 0$ as $B \to \infty$, we simply say that $\{\mathcal{C}_B\}$ achieves a delay that is linear in B—leaving implicit "with probability asymptotically equal to one."

A key parameter we shall be concerned with is

$$\beta \triangleq \frac{\log A}{B},$$

which we call the timing uncertainty per information bit.

Next, we define the asynchronous capacity per unit cost in the asymptotic regime where $B \to \infty$ while β is kept fixed.

Definition 4 (Asynchronous Capacity per Unit Cost)**.** $\boldsymbol{R}$ *is an achievable rate per unit cost at timing uncertainty per information bit β and delay exponent δ if there exists a sequence of codes $\{\mathcal{C}_B\}$, and a sequence of numbers $\{\varepsilon_B\}$ with $\varepsilon_B \overset{B\to\infty}{\longrightarrow} 0$, such that*

$$\mathbb{P}(\mathcal{E}|\mathcal{C}_B) \leq \varepsilon_B,$$

$$\limsup_{B\to\infty} \log(d(\mathcal{C}_B, \varepsilon_B))/B \leq \delta,$$

and

$$\liminf_{B\to\infty} \frac{B}{K(\mathcal{C}_B)} \geq \boldsymbol{R}.$$

The asynchronous capacity per unit cost, denoted by $\boldsymbol{C}(\beta, \delta)$, is the largest achievable rate per unit cost. In the important case when $\delta = 0$, we define $\boldsymbol{C}(\beta) \triangleq \boldsymbol{C}(\beta, 0)$.

Note that, in Definition 4, the codeword length n is a free parameter that can be optimized, just as for the synchronous capacity per unit cost (see the comment after [8, Definition 2]). The results in the next section characterize the capacity per unit cost for arbitrary β and δ. Similar to the synchronous case, the results simplify when there is a zero cost symbol, specifically when $\mathcal{X}$ contains $\star$ and $\star$ has zero cost.

For simplicity, for the rest of the paper we assume that the only possible zero cost symbol is $\star$—in particular, if $\star \notin \mathcal{X}$ then $\mathcal{X}$ contains only non-zero cost symbols. The other, arguably unnatural, cases can also be addressed by the arguments in this paper and are briefly discussed in the remark before the proof of Theorem 3 in Section IV.

[5]Hence, by definition we have

$$\mathbb{P}_m(\cdot) = \frac{1}{A} \sum_{t=1}^{A} \mathbb{P}_{m,t}(\cdot).$$

III. Results

Our first result gives the asynchronous capacity per unit cost when $\delta = 0$. It can be viewed as the asynchronous analogue of Theorem 2 in [8], which states that the synchronous capacity per unit cost is

$$\max_X \frac{I(X;Y)}{\mathbb{E}[k(X)]}\,. \tag{3}$$

As mentioned above, in stating our results we assume that all non-$\star$ symbols in $\mathcal{X}$ have positive cost, and that if $\star$ is in $\mathcal{X}$, then $\star$ has zero cost.

Theorem 1 (Asynchronous Capacity per Unit Cost: Sub-exponential Delay Constraint)**.** *The asynchronous capacity per unit cost at delay exponent $\delta = 0$ is given by*

$$\boldsymbol{C}(\beta) = \max_X \min\left\{\frac{I(X;Y)}{\mathbb{E}[k(X)]}, \frac{I(X;Y)+D(Y||Y_\star)}{\mathbb{E}[k(X)](1+\beta)}\right\}, \tag{4}$$

where X denotes the random input to the channel, Y the corresponding output, $Y_\star$ the random output of the channel when the idle symbol $\star$ is transmitted (i.e., $Y_\star \sim Q(\cdot|\star)$), $I(X;Y)$ the mutual information between X and Y, and $D(Y||Y_\star)$ the Kullback-Leibler distance between the distributions of Y and $Y_\star$.[6]

Furthermore, capacity can be achieved by codes whose delay grows linearly in B.[7]

The two terms in (4) reflect the two constraints on reliable communication. The first term corresponds to the standard constraint that the number of bits that can reliably be transmitted per channel use cannot exceed the input-output mutual information. This constraint applies when the channel is synchronous, hence also in the absence of synchrony.

The second term in (4) corresponds to the receiver's ability to determine the arrival time ν of the data. Indeed, even though the decoder is only required to produce a message estimate, because of the delay constraint, there is no loss in terms of capacity per unit cost to also require the decoder to produce an approximate estimate of the time when transmission begins—the delay constraint implies that the decoder can locate the sent message within a time window that is negligible compared to A. The quantity

$$I(X;Y) + D(Y||Y_\star) = D(XY||XY_\star),$$

where $D(XY||XY_\star)$ refers to the Kullback-Leibler distance between the joint distribution of (X,Y) and the (product) distribution of $(X,Y_\star)$, measures how difficult it is for the receiver to discern a data-carrying transmitted symbol from pure noise, and thus determines how difficult it is for the receiver to get the timing correct.

When the alphabet $\mathcal{X}$ contains a zero-cost symbol 0, the synchronous result (3) simplifies, and Theorem 3 in [8] says that the synchronous capacity per unit cost becomes

$$\max_{x\in\mathcal{X}} \frac{D(Y_x||Y_0)}{k(x)}, \tag{5}$$

[6] $Y_\star$ is interpreted as "pure noise."

[7] See comment after Definition 3.

an optimization over the input alphabet instead of over the set of all input distributions, where Y_x refers to the output distribution given that x is transmitted.

We find an analogous simplification in the asynchronous setting when $\star$ is in $\mathcal{X}$ and has zero cost:

Theorem 2 (Asynchronous Capacity per Unit Cost With Zero Cost Symbol: Sub-exponential Delay Constraint)**.** *If $\star$ is in $\mathcal{X}$ and has zero cost, the asynchronous capacity per unit cost at delay exponent $\delta = 0$ is given by*

$$\boldsymbol{C}(\beta) = \frac{1}{1+\beta}\max_{x\in\mathcal{X}} \frac{D(Y_x||Y_\star)}{k(x)}\,, \tag{6}$$

and capacity can be achieved by codes whose delay grows linearly with B.

Hence, a lack of synchronization multiplies the cost of sending one bit of information by $1+\beta$. An intuitive justification for this is as follows. Suppose there exists an optimal coding scheme that can both isolate and locate the sent message with high probability—as alluded to above, the ability to "locate" the message is a consequence of the decoder's delay constraint. Assuming that the delay is negligible, *i.e.*, the delay grows subexponentially with B, this allows us to consider message/location pairs as inducing a code of size

$$\approx 2^B A$$

used for communication across the *synchronous channel*. Hence, since $A = 2^{\beta B}$ we are effectively communicating

$$\approx \beta B + B = B(1+\beta)$$

bits reliably over the synchronous channel. Therefore, sending B bits of information at asynchronism level β is at least as costly as sending $B(1+\beta)$ bits over the synchronous channel. Flipping this reasoning around, the asynchronous channel effectively induces a codebook for message/location pairs where the location is encoded via *pulse position modulation* (PPM). From [8], optimal coding schemes are similar to PPM in that the codewords consist almost entirely of the zero cost symbol. This provides an intuitive justification for why $(1+\beta)k_{\text{sync}}$ is an achievable rate per unit cost.

Theorem 2 can be extended to the (continuous-valued) Gaussian channel, where the idle symbol $\star$ is the 0-symbol:

Theorem 3 (Asynchronous Capacity per Unit Cost for the Gaussian Channel: Sub-exponential Delay Constraint)**.** *The asynchronous capacity per unit cost for the Gaussian channel with variance $N_0/2$, quadratic cost function (i.e., $k(x) = x^2$), and delay exponent $\delta = 0$, is given by*

$$\boldsymbol{C}(\beta) = \frac{1}{1+\beta}\frac{\log e}{N_0}, \quad \beta \geq 0\,. \tag{7}$$

Theorem 1 can be extended to the case of a large delay constraint, *i.e.*, when $0 < \delta \leq \beta$. In this case, the formula for capacity is slightly different depending on whether $\star$ is in $\mathcal{X}$ or not, as stated in the following result.

Theorem 4 (Asynchronous Capacity per Unit Cost: Exponential Delay Constraint)**.** *The asynchronous capacity per unit cost at delay constraint δ, with $0 \leq \delta \leq \beta$, is given by:*

(a) *if $\star \in \mathcal{X}$ and $\star$ has zero cost then*

$$\boldsymbol{C}(\beta, \delta) = \boldsymbol{C}(\beta - \delta),$$

i.e., it is the same as the capacity per unit cost with delay exponent $\delta = 0$, but with asynchronism exponent β reduced to $\beta - \delta$;

(b) *if $\star$ is not in $\mathcal{X}$ and all non-$\star$ symbols have positive cost then*

$$\boldsymbol{C}(\beta, \delta) = \max_X \min \left\{ \frac{I(X;Y)}{\mathbb{E}[k(X)](1-\delta)}, \frac{I(X;Y) + D(Y||Y_\star)}{\mathbb{E}[k(X)](1+\beta-\delta)} \right\}. \quad (8)$$

The uniform distribution on ν in the model is not critical. The next result extends Theorem 1 to the case where ν is non-uniform. For a non-uniform distribution on ν, what is important turns out to be its "smallest" set of mass points that contains "most" of the probability.

Consider a general arrival time ν (defined over the positive integers), not necessarily bounded. For a given $\varepsilon > 0$, let $\mathcal{S}(\varepsilon)$ denote the smallest subset of the support of ν (*i.e.*, the set of n such that $\mathbb{P}(\nu = n) > 0$) whose probability is at least $1 - \varepsilon$. Hence, $\mathbb{P}(\nu \in \mathcal{S}(\varepsilon)) \geq 1 - \varepsilon$ by definition.

Theorem 5 (Asynchronous Capacity per Unit Cost With Non-uniform Arrival Time: Sub-exponential Delay Constraint)**.** *For a given sequence of arrival times $\{\nu_B\}_{B \geq 1}$, define*

$$\bar{\beta} = \inf_{\{\varepsilon_B\}} \limsup_{B \to \infty} \frac{\log(|\mathcal{S}(\varepsilon_B)|)}{B}, \quad (9)$$

where the infimum is with respect to all sequences $\{\varepsilon_B\}$ of nonnegative numbers such that $\lim_{B\to\infty} \varepsilon_B = 0$.

Then, the asynchronous capacity per unit cost at delay exponent 0 is given by

$$\boldsymbol{C}(\bar{\beta}) = \max_X \min \left\{ \frac{I(X;Y)}{\mathbb{E}[k(X)]}, \frac{I(X;Y) + D(Y||Y_\star)}{\mathbb{E}[k(X)](1+\bar{\beta})} \right\}.$$

Although the formula for $\bar{\beta}$ in (9) appears unwieldy, in many cases it can easily be evaluated. For example, in many cases, such as the uniform or geometric distributions, the formula reduces to the normalized entropy

$$\bar{\beta} = \lim_{B \to \infty} H(\nu_B)/B\,.$$

There are cases, however, where (9) doesn't reduce to the normalized entropy. For instance, consider the case when $\nu_B = 1$ with probability $1/2$, and $\nu_B = i$ with probability $(1/2)2^{-\beta B}$ for $i = 2, \ldots, 2^{\beta B} + 1$. Then, $\bar{\beta} = \beta$ and $H(\nu_B) = 1 + 0.5\beta B$, which yields

$$\bar{\beta} = 2 \lim_{B \to \infty} H(\nu_B)/B\,.$$

Asynchronous Capacity

The above results focus on characterizing the asynchronous capacity *per unit cost*. However, just as the synchronous capacity per unit cost result (3) immediately implies the standard (synchronous) capacity result[8]

$$C = \max_X I(X;Y)$$

by setting the cost function $k(\cdot) = 1$, Theorem 1 implies the asynchronous capacity result

$$C(\beta) = \max_X \min \left\{ I(X;Y); \frac{I(X;Y) + D(Y||Y_\star)}{1+\beta} \right\}, \quad (10)$$

the largest number of information bits per *transmitted symbol* that can be supported reliably by an asynchronous channel, as a function of β.

Instead of β, we may alternatively consider the asynchronism parameter $\alpha = (\log A)/n = \beta R$ introduced in [1], [7]. Using (10), we deduce that rate R is achievable if and only if, for some input X,

$$R \leq I(X;Y)$$

and

$$R \leq D(XY||XY_\star) - \alpha\,.$$

Hence, asynchronous capacity is alternatively given by

$$C(\alpha) = \max \Big\{ \max_{X: D(Y||Y_\star) \geq \alpha} I(X;Y); \max_{X: D(Y||Y_\star) \leq \alpha} D(XY||XY_\star) - \alpha \Big\}, \quad (11)$$

with the convention that the maximum evaluates to 0 if the set being optimized over is empty. Consider the second inner maximization in (11). Since $D(XY||XY_\star)$ is convex in X, and the set $\{X : D(Y||Y_\star) \leq \alpha\}$ is convex, the maximum is achieved for some extreme point of the set, *i.e.*, either for some X such that $D(Y||Y_\star) = \alpha$, or for a distribution X concentrated on a single point and such that $D(Y||Y_\star) < \alpha$. However, in the latter case we have

$$D(XY||XY_\star) - \alpha < 0$$

since $D(XY||XY_\star) = D(Y||Y_\star) < \alpha$. Thus, (11) reduces to

$$C(\alpha) = \max_{X: D(Y||Y_\star) \geq \alpha} I(X;Y)\,.$$

Although not explicit in the statement of Theorem 1, the proof of this theorem shows that $C(\alpha)$ can be achieved with codes whose delays are no larger than n. Summarizing the above discussion, we get:

Corollary. *The capacity at delay exponent $\delta = 0$, and with respect to asynchronism parameter $\alpha = (\log A)/n$, is given by*

$$C(\alpha) = \max_{X: D(Y||Y_\star) \geq \alpha} I(X;Y)\,.$$

Furthermore, capacity is achievable with codes whose delays are no larger than n.

A closely related problem is determining the capacity when rate is defined in terms of bits per *received symbol*. For this problem, we refer the reader to [7], [6], where capacity as a

[8]Information per symbol and information per unit cost are differentiated by lightface and boldface characters, respectively, as in [8].

function of α is studied, and where rate is defined with respect to the expected elapsed time between the instant information is available at the transmitter and the instant it is decoded.

IV. Proofs of Results

We use $\mathcal{P}^{\mathcal{X}}$ to denote the set of distributions over the finite alphabet $\mathcal{X}$. Recall that the type of a string $x^n \in \mathcal{X}^n$, denoted by $\hat{P}_{x^n}$, is the probability distribution over $\mathcal{X}$ that assigns, to each $a \in \mathcal{X}$, the number of occurrences of a within x^n divided by n [4, Chapter 1.2]. For instance, if $x^3 = 010$, then $\hat{P}_{x^3}(0) = 2/3$ and $\hat{P}_{x^3}(1) = 1/3$. The joint type $\hat{P}_{x^n,y^n}$ induced by a pair of strings $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is defined similarly. The set of strings of length n that have type P is denoted by $\mathcal{T}_P$, and is called the "type class of P." The set of all types over $\mathcal{X}$ of strings of length n is denoted by $\mathcal{P}_n^{\mathcal{X}}$.

Given a string $x^n \in \mathcal{X}^n$ and a conditional probability distribution $W = \{W(y|x),\ (x,y) \in \mathcal{X} \times \mathcal{Y}\}$, the set of strings y^n that have conditional type W given x^n is denoted by $\mathcal{T}_W(x^n)$, *i.e.*,

$$\mathcal{T}_W(x^n) \triangleq \{y^n \in \mathcal{Y}^n : \hat{P}_{x^n,y^n} = \hat{P}_{x^n} W\}\,.$$

Finally, we use the standard "big-O" Landau notation to characterize growth rates (see, e.g., [2, Chapter 3]), and use $\mathrm{poly}(\cdot)$ to denote a function that does not grow or decay faster than polynomially in its argument.

The following two standard results on types are often used in the analysis:

Fact 1 ([4, Lemma 2.2])**.**

$$|\mathcal{P}_n^{\mathcal{X}}| = \mathrm{poly}(n)\,.$$

Fact 2 ([4, Lemma 2.6])**.** *If X^n is independent and identically distributed (i.i.d.) according to $X_1 \sim P_1$, then*

$$\mathrm{poly}(n) e^{-nD(X_2\|X_1)} \le \mathbb{P}(X^n \in \mathcal{T}_{P_2}) \le e^{-nD(X_2\|X_1)}$$

for any $X_2 \sim P_2 \in \mathcal{P}_n^{\mathcal{X}}$.

Achievability of Theorem 1: We first show the existence of a random code that achieves the asynchronous capacity per unit cost when the latter is computed with respect to average error probability. A standard expurgation argument then shows the existence of a deterministic code achieving the same (asymptotic) performance as the random code, but now with respect to maximum error probability.

Fix some arbitrary distribution P on $\mathcal{X}$. Let X be the input having that distribution, and let Y be the corresponding output, *i.e.*, $(X,Y) \sim P(\cdot)Q(\cdot|\cdot)$.

Given B bits of information to be transmitted, the codebook $\mathcal{C}$ is randomly generated as follows. For each message $m \in \{1, 2, \ldots, 2^B\}$, randomly generate a length n sequence x^n i.i.d. according to P. If x^n belongs to the "constant composition" set[9]

$$\mathcal{A} = \{x^n : ||\hat{P}_{x^n} - P|| \le 1/\log n\}\,, \tag{12}$$

we let $c^n(m) = x^n$. Otherwise, we repeat the procedure until we generate a sequence sufficiently close to P. From Chebyshev's inequality, for a fixed m, it is very unlikely that any repetition will be required to generate $c^n(m)$, *i.e.*,

$$P^n(\mathcal{A}) \to 1 \quad \text{as} \quad n \to \infty, \tag{13}$$

where P^n denotes the order n product distribution of P.

The obtained codebook is thus essentially of constant composition, *i.e.*, each symbol appears roughly the same number of times across codewords. Moreover, by construction all codewords in the random ensemble have cost

$$n\mathbb{E}[k(X)](1 + o(1))$$

as $n \to \infty$.

The sequential typicality decoder operates as follows. At time t, for each $m \in \{1, 2, \ldots, 2^B\}$, it computes the empirical distributions

$$\hat{P}_{c^n(m), y_{t-n+1}^t}(\cdot,\cdot)$$

induced by $c^n(m)$ and the n output symbols y_{t-n+1}^t. If there is a unique message m for which

$$||\hat{P}_{c^n(m), y_{t-n+1}^t}(\cdot,\cdot) - P(\cdot)Q(\cdot|\cdot)|| \le 2/\log n,$$

the decoder stops and declares that message m was sent. If more than one codeword is typical, the decoder stops and declares one of the corresponding messages uniformly at random.[10] If no codeword is typical at time t, the decoder moves one step ahead and repeats the procedure based on Y_{t-n+2}^{t+1}. If the decoder reaches time $A+n-1$ and no codeword is typical, then it declares a randomly and uniformly chosen message.

We first compute the error probability averaged over codebooks and messages. Suppose message m is transmitted. The error event that the decoder declares some specific message $m' \neq m$ can be decomposed as[11]

$$\{m \to m'\} = \mathcal{E}_1 \cup \mathcal{E}_2\,, \tag{14}$$

where the error events $\mathcal{E}_1$ and $\mathcal{E}_2$ are defined as

- $\mathcal{E}_1$: the decoder stops at a time t between ν and $\nu+2n-2$ (including ν and $\nu+2n-2$), and declares m';
- $\mathcal{E}_2$: the decoder stops either at a time t before ν or from $\nu+2n-1$ onwards, and declares m'.

For the error event $\mathcal{E}_1$, for some $0 \le k \le n-1$ the first or the last k symbols of Y^n are generated by noise, and the remaining $n-k$ symbols are generated by the sent codeword $C^n(m)$.[12] The probability that such a Y^n together with $C^n(m')$ yields an empirical distribution J that is jointly typical with $P(\cdot)Q(\cdot|\cdot)$, that is,

$$||J(\cdot,\cdot) - P(\cdot)Q(\cdot|\cdot)|| \le 2/\log n\,, \tag{15}$$

[9] $||\cdot||$ refers to the L_1-norm.

[10] The notion of typicality we use is often referred to as "strong typicality" in the literature.

[11] Notice that the decoder outputs a message with probability one by time $A+n-1$.

[12] We use a capital letter for $C^n(m)$ since codewords are randomly generated.

is upper bounded as

$$\begin{aligned}
&\mathbb{P}_m(\hat{P}_{C^n(m'),Y^n} = J) \\
&= \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}_m(Y^n = y^n) \sum_{x^n : \hat{P}_{x^n,y^n} = J} \mathbb{P}_m(X^n = x^n) \\
&\leq \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}_m(Y^n = y^n) \sum_{x^n : \hat{P}_{x^n,y^n} = J} 2^{-n(H(J_{\mathcal{X}}) + D(J_{\mathcal{X}}||P) - \varepsilon)} \\
&\leq \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}_m(Y^n = y^n) 2^{-n(H(J_{\mathcal{X}}) - \varepsilon)} |\{x^n : \hat{P}_{x^n,y^n} = J\}| \\
&\leq \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}_m(Y^n = y^n) 2^{-n(H(J_{\mathcal{X}}) - \varepsilon)} 2^{nH(J_{\mathcal{X}|\mathcal{Y}})} \\
&\leq 2^{-n(I(J) - \varepsilon)} \\
&\leq 2^{-n(I(X;Y) - 2\varepsilon)}
\end{aligned} \tag{16}$$

for any $\varepsilon > 0$ and all n large enough, where $H(J_{\mathcal{X}})$ denotes the entropy of the left marginal of J, where

$$H(J_{\mathcal{X}|\mathcal{Y}}) \triangleq -\sum_{b \in \mathcal{Y}} J_{\mathcal{Y}}(b) \sum_{a \in \mathcal{X}} J_{\mathcal{X}|\mathcal{Y}}(a|b) \log J_{\mathcal{X}|\mathcal{Y}}(a|b),$$

and where $I(J)$ denotes the mutual information induced by J.

The first equality in (16) follows from the independence of $C^n(m')$ and Y^n, since Y^n corresponds to the output of $C^n(m)$. For the first inequality, note that if the codewords were randomly generated with each component of each codeword i.i.d. according to P, we could deduce from [3, Theorem 11.1.2, p. 349] that

$$P^n(X^n = x^n) = 2^{-n(H(J_{\mathcal{X}}) + D(J_{\mathcal{X}}||P))}.$$

The actual (non-i.i.d) codeword distribution is the i.i.d. distribution, conditioned on the constant composition event (12). Therefore, we have

$$\mathbb{P}_m(X^n = x^n) = \begin{cases} \frac{P^n(X^n = x^n)}{P^n(\mathcal{A})}) & x^n \in \mathcal{A} \\ 0 & \text{otherwise,} \end{cases}$$

and from (13) we get

$$\mathbb{P}_m(X^n = x^n) = 2^{-n(H(J_{\mathcal{X}}) + D(J_{\mathcal{X}}||P))}(1 + o(1))$$

as $n \to \infty$, uniformly over the set $\mathcal{A}$. This justifies the first inequality in (16). The second inequality in (16) follows from the non-negativity of the Kullback-Leibler distance. The third inequality in (16) follows from [4, Lemma 2.5, p. 31]. The fourth inequality holds since $H(J_{\mathcal{X}}) - H(J_{\mathcal{X}|\mathcal{Y}}) = I(J)$, and by upperbounding the sum of the probabilities by one. Finally, the fifth inequality in (16) holds for any $\varepsilon > 0$ and all n large enough since, by assumption, J is close to PQ (see (15)).

From (16), by taking a union bound over all empirical distributions J that are jointly typical with PQ ($\mathrm{poly}(n)$ by Fact 1) and over all the (less than $2n$) times involved in $\mathcal{E}_1$, we obtain the upper bound

$$\mathbb{P}_m(\mathcal{E}_1) \leq 2^{-n(I(X;Y) - 3\varepsilon)} \tag{17}$$

for all n large enough.

For the second error event $\mathcal{E}_2$, pure noise produces some output Y^n that is jointly typical with $C^n(m')$. The probability that a noise generated Y^n together with $C^n(m')$ yields an empirical type J is upper bounded by

$$2^{-nD(J||XY_\star)}$$

by [4, Lemma 1.2.6]—recall that $D(J||XY_\star)$ refers to the Kullback-Leibler distance between, on the one hand, the joint distribution J, and on the other hand, the product of the distributions of X and $Y_\star$. Hence, by taking a union bound over all typical J's that satisfy (15) ($\mathrm{poly}(n)$ of them by Fact 1), and by using the continuity of the Kullback-Leibler distance, [13] the probability that a noise generated Y^n is typical with $C^n(m')$ is upper bounded by

$$2^{-n(D(XY||XY_\star) - \varepsilon)} = 2^{-n(I(X;Y) + D(Y||Y_\star) - \varepsilon)},$$

for any $\varepsilon > 0$ and all n large enough. Finally, by taking a union bound over all (less than A) times where noise could produce such an output, we get

$$\mathbb{P}_m(\mathcal{E}_2) \leq A \cdot 2^{-n(I(X;Y) + D(Y||Y_\star) - \varepsilon)}, \tag{18}$$

for any $\varepsilon > 0$ and all n large enough.

Combining (14), (17), and (18), we get

$$\begin{aligned}
\mathbb{P}_m(m \to m') &= \mathbb{P}_m(\mathcal{E}_1) + \mathbb{P}_m(\mathcal{E}_2) \\
&\leq 2^{-n(I(X;Y) - 3\varepsilon)} \\
&\quad + A \cdot 2^{-n(I(X;Y) + D(Y||Y_\star) - \varepsilon)},
\end{aligned}$$

for any $\varepsilon > 0$ and all n large enough.

Hence, by taking a union bound over all possible wrong messages, we obtain that for any $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}_m(\mathcal{E}) \leq 2^B \Big(&2^{-n(I(X;Y) - 3\varepsilon)} \\
&+ A \cdot 2^{-n(I(X;Y) + D(Y||Y_\star) - \varepsilon)} \Big),
\end{aligned}$$

for n large enough and all m. Since the above bound is valid for a randomly generated code, we deduce that

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}(\bar{\mathbb{P}}(\mathcal{E}|\mathcal{C})) &= \mathbb{P}_m(\mathcal{E}) \\
&\leq 2^B \Big(2^{-n(I(X;Y) - 3\varepsilon)} \\
&\quad + A \cdot 2^{-n(I(X;Y) + D(Y||Y_\star) - \varepsilon)} \Big) \\
&\triangleq \varepsilon_1(n),
\end{aligned} \tag{19}$$

where $\bar{\mathbb{P}}(\mathcal{E}|\mathcal{C})$ denotes the error probability of code $\mathcal{C}$ averaged over the messages.

We now turn to the delay of the code. Suppose message m is transmitted with a specific (non-random) codeword $c^n(m)$ that belongs to the set $\mathcal{A}$. If event

$$\{\tau \geq \nu + n\}$$

happens, then necessarily $Y_\nu^{\nu+n-1}$ isn't typical with $c^n(m)$. By Chebyshev's inequality, the probability of the latter event tends to zero as $n \to \infty$, hence

$$\mathbb{P}_m(\tau \leq \nu + n) \geq 1 - \varepsilon_2(n),$$

[13] Technically, the divergence is not continuous if, for example, both distributions are 0 at the same point. However, at points of discontinuity, the discontinuity can only help since the divergence becomes infinite, and it is easily seen that the corresponding error event has zero probability.

where $\varepsilon_2(n)$ is a function that tends to zero as $n \to \infty$. Since the above inequality holds for any specific codeword that belongs to $\mathcal{A}$, we get

$$d(\mathcal{C}, \varepsilon_2(n)) \leq n \tag{20}$$

for any code $\mathcal{C}$ whose codewords belong to $\mathcal{A}$.

The proof can now be concluded. From inequality (19), there exists a specific code $\mathcal{C} \subset \mathcal{A}$ whose error probability, averaged over messages, is less than $\varepsilon_1(n)$. Removing the half of the codewords with the highest error probability, we end up with a set $\mathcal{C}'$ of 2^{B-1} codewords whose maximum error probability $\mathbb{P}(\mathcal{E})$ satisfies

$$\mathbb{P}(\mathcal{E}) \leq 2\varepsilon_1(n)\,, \tag{21}$$

and whose delay satisfies

$$d(\mathcal{C}', \varepsilon_2(n)) \leq n$$

by the previous argument.

Now, fix the ratio B/n, thereby imposing a delay linear in B, and substitute $A = 2^{\beta B}$ in the definition of $\varepsilon_1(n)$ (see (19)). Then, $\mathbb{P}(\mathcal{E})$ goes to zero as $B \to \infty$ whenever

$$\frac{B}{n} < \min\left\{ I(X;Y), \frac{I(X;Y) + D(Y||Y_\star)}{1+\beta} \right\}. \tag{22}$$

Recall that, by construction, all the codewords have cost $n\mathbb{E}[k(X)](1+o(1))$ as $n \to \infty$. Hence, for any $\eta > 0$ and all n large enough,

$$k(\mathcal{C}') \leq n\mathbb{E}[k(X)](1+\eta)\,. \tag{23}$$

Condition (22) is thus implied by condition

$$\frac{B}{K(\mathcal{C}')} < \min\left\{ \frac{I(X;Y)}{(1+\eta)\mathbb{E}[k(X)]}, \frac{I(X;Y) + D(Y||Y_\star)}{\mathbb{E}[k(X)](1+\eta)(1+\beta)} \right\}. \tag{24}$$

Maximizing over all input distributions, and using the fact that $\eta > 0$ can be chosen arbitrarily, proves that the right-hand side of (4) is asymptotically achieved by non-random codes with delay at most n, which grows linearly with B. ∎

Remark. *From* (24) *it follows that whenever there exists some input X such that $I(X;Y) > 0$ while $\mathbb{E}[k(X)] = 0$, and thus $\mathcal{X}$ contains more than one zero cost symbol, the asynchronous capacity per unit cost is infinite, i.e., $\boldsymbol{C}(\beta) = \infty$, for any $\beta \geq 0$.*

Achievability of Theorem 4: The achievability scheme for Theorem 4 is similar to the achievability scheme used to prove Theorem 1 except that we distinguish the cases $\star \in \mathcal{X}$ and $\star \notin \mathcal{X}$.

(a) $\star \in \mathcal{X}$: The main change is that now the transmitter does not start transmitting at time ν. Instead, the transmitter only starts transmitting at the first multiple of $2^{\delta B}$ larger than ν, so that now σ takes values over multiples of $2^{\delta B}$. Such a transmission scheme reduces the receiver's uncertainty about σ from uniformly over $2^{\beta B}$ time slots to (essentially) uniformly over only $2^{(\beta-\delta)B}$ time slots.

One proves that $\boldsymbol{C}(\beta-\delta)$ is achievable with delay $O(2^{\delta B})$ by repeating the arguments for the achievability of Theorem 1. The random codebook is constructed so that each codeword satisfies the constant composition property. The blocklength n is still chosen to be $O(B)$ so that, in contrast with the achievability of Theorem 1, where delay and blocklength are the same, now the blocklength is exponentially smaller than the delay.

The rest of the analysis is essentially unchanged. Since the codewords are constructed in the same way, the cost is unchanged, and the probability of error analysis is the same, except that A is replaced by $A/2^{\delta B}$ because now the transmission timing allows the decoder to only consider $A/2^{\delta B}$ time slots instead of all A time slots. Therefore, β is replaced by $\beta - \delta$, completing the proof.

(b) $\star \notin \mathcal{X}$: The main change is that the transmitter uses the freedom in the choice of σ to communicate part of the information through timing; $B(1-\delta)$ information bits are contained in each codeword and $B\delta$ information bits are conveyed via timing. To achieve this, we use a space-time code.

The transmitter generates $2^{B(1-\delta)}$ random codewords in the same way as in the achievability proof of Theorem 1 to obtain a codebook

$$\{c^n(s) \quad \text{with} \quad 1 \leq s \leq 2^{(1-\delta)B}\}\,.$$

Label each of the 2^B messages with one of the $2^{(1-\delta)B} \times 2^{\delta B}$ pairs of integer indices (s, j), *i.e.*, the message set is given by

$$\{m(s,j) \quad \text{with} \quad 1 \leq s \leq 2^{(1-\delta)B}, 1 \leq j \leq 2^{\delta B}\}\,.$$

(For simplicity we assume that $2^{B(1-\delta)}$ and $2^{\delta B}$ are integers.) For any (space) index $s \in \{1, 2, \ldots, 2^{(1-\delta)B}\}$, the set of messages

$$\{m(s,j), 1 \leq j \leq 2^{\delta B}\}$$

is associated to codeword $c^n(s)$.

Transmission always starts at a time that is a multiple of n. Suppose message m arrives at time ν and that $m = m(\bar{s}, \bar{j})$. The transmitter first computes the "offset"

$$O = \bar{j} - \left\lceil \frac{\nu}{n} \right\rceil \mod 2^{\delta B}.$$

The transmitter then starts sending codeword $c^n(\bar{s})$ at time

$$\sigma(\nu, m) = \left(\left\lceil \frac{\nu}{n} \right\rceil + O \right) n. \tag{25}$$

The receiver uses a sequential typicality decoder to find the transmitted codeword as in the proof of the achievability part of Theorem 1—since transmission times are restricted to be multiples of n, the sequential typicality decoder can be restricted to multiples of n.

Suppose codeword $\hat{s}$ is found to be typical at time t. The receiver then computes the estimate $\hat{\sigma}$ for σ given by

$$\hat{\sigma} = t - n + 1$$

and finds the index $\hat{j} \in \{1, 2, \ldots, 2^{(1-\delta)B}\}$ such that

$$\hat{j} = \frac{\hat{\sigma}}{n} \mod 2^{\delta B}\,.$$

The receiver then declares $\hat{m} = m(\hat{s}, \hat{j})$.

The rest of the analysis is essentially unchanged. Since the codewords are constructed in the same way, the cost is

unchanged, and the probability of error analysis is the same, except that 2^B is replaced by $2^{B(1-\delta)}$ because the transmission timing allows the decoder to only consider $2^{B(1-\delta)}$ codewords instead of 2^B codewords. ∎

Achievability of Theorem 5: To prove the achievability part of Theorem 5, one applies essentially the same arguments as for the achievability of Theorem 1. The transmitter's strategy is unchanged, *i.e.*, $\sigma = \nu$, and a random codebook satisfying the constant composition property is used to encode the messages. At the receiver, we need a suitable analog of the set $\{1, 2, \ldots, A\}$ of time slots to consider. A natural choice is to pick a sequence of nonnegative numbers $\{\varepsilon_B\}$ such that $\varepsilon_B \stackrel{B\to\infty}{\longrightarrow} 0$, and, for each B, consider the "typical" set $\mathcal{S}(\varepsilon_B)$ whose probability, under the arrival time distribution, is at least $1-\varepsilon_B$ by definition. The receiver operates just as before, *i.e.*, using a sequential typicality decoder, but only over the set of times in $\mathcal{S}(\varepsilon_B)$.

Since the codewords are constructed in the same way, the cost of the codebook is unchanged. The probability of error and delay analysis now breaks into two cases: $\nu \in \mathcal{S}(\varepsilon_B)$ and $\nu \notin \mathcal{S}(\varepsilon_B)$. The case $\nu \in \mathcal{S}(\varepsilon_B)$ is handled as previously, except that A is replaced by $|\mathcal{S}(\varepsilon_B)|$. When $\nu \notin \mathcal{S}(\varepsilon_B)$, we make the worst-case assumption that the message is wrongly decoded and that the delay is infinite. We can afford to do this because $\mathbb{P}(\nu \notin \mathcal{S}(\varepsilon_B)) \stackrel{B\to\infty}{\longrightarrow} 0$ by definition. Hence, the event $\{\nu \notin \mathcal{S}(\varepsilon_B)\}$ has a vanishing effect on the probability of error and the delay. Optimizing over the choice of sequence $\{\varepsilon_B\}$ completes the proof. ∎

Converses of Theorems 1 and 4:

Assume that $\{\mathcal{C}_B\}$ achieves a rate per unit cost $\boldsymbol{R} > 0$ at timing uncertainty per information bit β and delay exponent δ with $0 \le \delta \le \beta$. Recall that the delay constraint means that

$$\limsup_{B\to\infty} \frac{\log d_B(\mathcal{C}_B, \varepsilon_B)}{B} = \delta \tag{26}$$

for some sequence of non-negative numbers $\varepsilon_B \to 0$ as $B \to \infty$. To establish the converses, we use the following concept of "extended codewords." To shorten notation, for the rest of the proof we use d_B instead of $d(\mathcal{C}_B, \varepsilon_B)$.

Extended codewords: An extended codeword for a given message m consists of the sequence of symbols that are transmitted from time ν until time $\nu + d_B - 1$. Hence, for $\nu + d_B - 1 \ge \sigma + n$, the codeword corresponding to message m consists of $\star$'s from time ν until time $\sigma - 1$, followed by $c^n(m)$, followed by $\star$'s until time $\nu + d_B - 1$. Instead, if $\nu + d_B \le \sigma + n$, the codeword corresponding to message m consists of $\star$'s from time ν until time $\sigma - 1$, followed by the first $\nu + d_B - \sigma$ symbols of $c^n(m)$. The cost of the extended codeword, which we simply denote by $c(m)$, is defined to be the same as the cost of $c^n(m)$.

From now on, codewords always refer to extended codewords, and codebooks always refer to sets of extended codewords.

To establish the theorems, we show that for any $\eta > 0$ and all B large enough, $\boldsymbol{R}$ and β satisfy

$$\boldsymbol{R}\mathbb{E}[k(X)] \le I(X;Y)(1+\eta) \tag{27}$$

if $\star \in \mathcal{X}$ and $\star$ has zero cost, or

$$\boldsymbol{R}\mathbb{E}[k(X)] \le \frac{I(X;Y)}{1-\delta}(1+\eta) \tag{28}$$

if $\star \notin \mathcal{X}$ and all non-$\star$ symbols have positive cost. In either case, we also show that

$$\boldsymbol{R}\mathbb{E}[k(X)](1+\beta-\delta-\eta) \le D(XY||XY_\star), \tag{29}$$

where $X \sim P_B$, and where P_B denotes the distribution of the type class of $\mathcal{C}_B$ which contains the most elements. This type class is denoted by $\mathcal{C}'_B$ in the sequel.

An important observation used to prove (27) and (29) is that because $\boldsymbol{R}$ can be assumed to be strictly positive (or there is nothing to prove), the set of non-$\star$ symbols of each codeword in $\mathcal{C}_B$ has at most $O(B)$ elements.

(Note that P_B may vary as a function of ν. However, for ease of exposition, we assume that P_B is the same for all ν. This assumption is without loss of generality, because we can group the ν's together based on their associated P_B, and as will become apparent from the analysis, our arguments can be applied to each group separately. Since $A = 2^{\beta B}$, for subsets containing at least $A2^{-\sqrt{B}}$ ν's, our arguments will be valid since $\liminf_{B\to\infty}(1/B)\log(A2^{-\sqrt{B}}) = \beta$. For P_B's associated with fewer than this many ν's, since there are only a polynomial number of P_B's, the probability of ν having any such P_B is $o(1)$.)

A. Proof of (27) and (28)

The intuition for these inequalities is that an asynchronous code must also be good for the synchronous channel, and hence a suitable notion of rate is bounded by the synchronous channel capacity. Formally, $\mathcal{C}'_B$ is clearly a good code for the synchronous channel, *i.e.*, if we reveal ν to the receiver and decoding happens at time $\nu + d_B$, it is possible to achieve an error probability bounded away from 1 whenever B is large enough. From the strong converse for synchronous communication (see, *e.g.*, [4, Corollary 6.4, p. 87]) it follows that when $\star \in \mathcal{X}$ and $\star$ has zero cost, for any $\eta > 0$,

$$\frac{\log|\mathcal{C}'_B|}{d_B} \le I(X;Y)(1+\eta/2) \tag{30}$$

for all B large enough. Similarly, when $\star \notin \mathcal{X}$, for any $\eta > 0$,

$$\frac{\log|\mathcal{C}'_B|}{n} \le I(X;Y)(1+\eta/2) + \frac{\delta B}{n} \tag{31}$$

for all B large enough, where n denotes the number of non-$\star$ symbols in each codeword. This can be seen by observing that the codewords can be classified according to the value of σ, and for a given σ, only a rate of $I(X;Y)(1+\eta/2)$ can be supported. Because of the delay constraint, only $2^{\delta B}$ choices of σ are possible.

Now, since the number of non-$\star$ symbols in any codeword is $O(B)$, the number of possible types P_B grows no faster than polynomially with B. To see this, note that there are $|\mathcal{X}|$ input symbols, and we have $O(B)$ choices for the probability assigned to each non-$\star$ symbol. Since there is at most one zero cost symbol (namely, the $\star$ symbol), P_B is completely

determined by the number of occurrences of the non-$\star$ symbols. Thus, there are only a total of $O(B^{|\mathcal{X}|})$ possible types P_B satisfying the constraint of having $O(B)$ non-$\star$ symbols. This implies that

$$\frac{\log|\mathcal{C}'_B|}{d_B} = \frac{\log|\mathcal{C}_B|}{d_B}(1-o(1))$$

when $\star \in \mathcal{X}$, and similarly for the case when $\star \notin \mathcal{X}$. Combining this with (30) and (31), we obtain

$$\frac{\log|\mathcal{C}_B|}{d_B} \le I(X;Y)(1+\eta)$$

when $\star \in \mathcal{X}$, and

$$\frac{\log|\mathcal{C}_B|}{n} \le I(X;Y)(1+\eta) + \delta B/n$$

when $\star \notin \mathcal{X}$. Note that $\log|\mathcal{C}_B| = B$ by definition. Thus, by multiplying and dividing the left-hand sides of the above inequalities by $K(\mathcal{C}_B)$, and by noting that $K(\mathcal{C}'_B) \le K(\mathcal{C}_B)$ by the definition of the cost of a code (see Definition 2 and recall that by definition, the extended codeword for message m has the same cost as $c^n(m)$), the above inequalities become

$$\frac{K(\mathcal{C}'_B)}{d_B}\boldsymbol{R} \le I(X;Y)(1+\eta)$$

and

$$\frac{K(\mathcal{C}'_B)}{n}\boldsymbol{R} \le \frac{I(X;Y)}{1-\delta}(1+\eta)\,.$$

Since $K(\mathcal{C}'_B) = d_B\mathbb{E}[k(X)]$ when $\star \in \mathcal{X}$ and $K(\mathcal{C}'_B) = n\mathbb{E}[k(X)]$ when $\star \notin \mathcal{X}$, inequalities (27) and (28) follow. Hence, if (27) or (28), as appropriate, doesn't hold, then the maximal error probability tends to one.

B. Proof of (29)

We show that if inequality (29) is reversed, then a decoder that satisfies the delay constraint has an average error over messages that tends to one. To prove this, we introduce the concepts of "effective output process" and "augmented decoder."

Effective output process: The "effective" output process is the random output process "viewed" by the sequential decoder, *i.e.*, it is generated as if there were pure noise after the transmission of the *extended* codeword. Specifically, the distribution of the effective output process is as follows. The Y_i's for[14]

$$i \in \{1,\dots,\nu-1\} \cup \{\nu+d_B,\dots,A_n+n-1\}$$

are i.i.d. according to $Q_\star$, whereas the block

$$Y_\nu, Y_{\nu+1}, \dots, Y_{\nu+d_B-1}$$

is distributed according to

$$\frac{1}{|\mathcal{C}'_B|}\sum_m Q(\cdot|c(m))\,,$$

the output distribution given that a *randomly selected* (extended) codeword from $\mathcal{C}'_B$ has been transmitted. With a slight abuse of notation, in the remainder of the proof we use $Y_1, Y_2, \dots, Y_{A+n-1}$ to denote the effective output process.

[14]Notice that because of (27), d_B is a strictly positive quantity.

Fig. 2. Parsing of the entire received sequence of size $A+n-1$ into r_B blocks of length d_B, one of which is generated by the sent message, while the others are generated by pure noise.

Augmented decoder: An augmented decoder is a decoder which is revealed the complete effective output sequence and, in addition, is informed that the message was sent in one of

$$r_B \triangleq \left\lfloor \frac{A+n-1-\nu \bmod d_B}{d_B} \right\rfloor \tag{32}$$

consecutive (disjoint) blocks of duration d_B, as shown in Fig. 2. Note that[15]

$$r_B \doteq 2^{B(\beta-\delta)}\,. \tag{33}$$

An augmented decoder, in addition to outputting a message, also outputs an estimate of the block of size d_B corresponding to the time interval during which the message was sent.

Suppose the decoder of $\mathcal{C}'_B$ achieves (maximum) communication delay less than d_B with probability equal to $1-\tilde{\varepsilon}_B$. Further, suppose it can output the correct message with maximum error probability ε_B. Hence, the corresponding augmented decoder can both output the block of size d_B which corresponds to the actual transmission period, and output the correct message, with maximum error probability at most $\varepsilon_B + \tilde{\varepsilon}_B$. We now show that if (29) doesn't hold, then with probability approaching one, pure noise will produce many output blocks that look as if they were generated by some codeword. This implies that $\varepsilon_B + \tilde{\varepsilon}_B \to 1$. Therefore, if the delay constraint is satisfied with $\tilde{\varepsilon}_B \to 0$, then $\varepsilon_B \to 1$. Hence, if the decoder of $\mathcal{C}'_B$ achieves (maximum) communication delay less than d_B with probability tending to one, its error probability will tend to one whenever (29) doesn't hold.

To develop some intuition for proving (29), we first consider the simpler setting where there is only a single message. We then generalize to the multiple message case to obtain (29).

1) Single message: Suppose there is only one codeword to be transmitted. The augmented decoder's only task is thus to output the block of size d_B that corresponds to the period when $c(m)$ was sent.

For this specific setting, we show that if β is sufficiently large, the decoder will not be able to perform the task reliably, because the noise is likely to produce several blocks that look as though they were generated by $c(m)$. More precisely, we show that the augmented decoder has a large probability of error (asymptotically equal to one) whenever for some $\eta > 0$ and all B large enough,

$$B(\beta-\delta-\eta) > d_B D(XY||XY_\star)\,. \tag{34}$$

[15]We use the notation $f(B) \doteq g(B)$ whenever the functions f and g are exponentially equal, *i.e.*, if

$$\lim_{B\to\infty}\frac{1}{B}\log f(B) = \lim_{B\to\infty}\frac{1}{B}\log g(B)\,.$$

Let $\bar{c}(m)$ denote the extended codeword $c(m)$ without zero-cost symbols and let $\bar{Y}(m)$ be its corresponding output. For instance, if the extended codeword is $c(m) = 1, 2, \star, 2, \star$ and its corresponding random output vector $Y(m)$ takes value $2, 2, 1, \star, 1$ then $\bar{c}(m) = 1, 2, 2$ and $\bar{Y}(m) = 2, 2, \star$. Further, let $\hat{Q}$ be the empirical distribution of $\bar{Y}(m)$ conditioned on $\bar{c}(m)$, *i.e.*, $\hat{Q}$ satisfies

$$\hat{P}_{\bar{c}(m),\bar{Y}(m)}(x, y) = \bar{P}_B(x)\hat{Q}(y|x),$$

where $\bar{P}_B$ denotes the empirical distribution of $\bar{c}(m)$.

The above restriction to the non-$\star$ symbols allows us to treat the various possible delays—linear in B, subexponential in B, and exponential in B—in a unified way. Had we been interested only in the linear case, the argument would also hold without the restriction to non-$\star$ symbols.

For a given fixed conditional probability distribution $\tilde{Q}$, denote by $Z(m, \tilde{Q})$ the binomial random variable which represents the number of pure noise blocks, out of $r_B - 1$ of them, whose conditional empirical distribution with respect to the non-$\star$ symbols of $\bar{c}(m)$ is $\tilde{Q}$. Then the error probability of the augmented decoder can be lower bounded as

$$\begin{aligned}&\mathbb{P}_m(\mathcal{E})\\ &\geq \sum_{\{\tilde{Q}:\tilde{Q}\approx Q\}} \mathbb{P}_m(\mathcal{E}|\hat{Q} = \tilde{Q}) \times \mathbb{P}_m(\bar{Y}(m) \in \mathcal{T}_{\tilde{Q}}(\bar{c}(m))),\end{aligned} \quad (35)$$

where the $\tilde{Q}$'s in the summation are conditional distributions that are close to the actual channel Q. Specifically, $\tilde{Q}(\cdot|x)$ is such that

$$||\tilde{Q}(\cdot|x) - Q(\cdot|x)|| \leq 1/\log B \quad (36)$$

for any symbol $x \neq \star$ that appears appears in $\bar{c}$ at least $\sqrt{B}$ times. And for any x that appears in $\bar{c}(m)$ less than $\sqrt{B}$ times, $\tilde{Q}(\cdot|x)$ is arbitrary.

Now, conditioned on $\{\hat{Q} = \tilde{Q}\}$, there are $Z(m, \tilde{Q})$ pure noise blocks which look statistically identical to the block corresponding to the sent codeword, because the empirical conditional distribution of (the non-$\star$ codeword symbol positions of) each block is a sufficient statistic for estimating the position of the sent codeword. Hence, the augmented decoder fails with probability at least

$$\mathbb{E}\left(\frac{Z(m, \tilde{Q})}{Z(m, \tilde{Q}) + 1}\right).$$

Therefore, from (35),

$$\begin{aligned}&\mathbb{P}_m(\mathcal{E})\\ &\geq \sum_{\{\tilde{Q}:\tilde{Q}\approx Q\}} \mathbb{E}\left(\frac{Z(m, \tilde{Q})}{Z(m, \tilde{Q}) + 1}\right) \mathbb{P}_m(\bar{Y}(m) \in \mathcal{T}_{\tilde{Q}}(\bar{c}(m))).\end{aligned} \quad (37)$$

From Fact 2, the probability that one single pure noise block induces the joint type $\bar{P}_B\tilde{Q}$ with $\bar{c}(m)$ is

$$\doteq 2^{-\bar{d}_B D(\bar{X}\tilde{Y}||\bar{X}Y_\star)} \doteq 2^{-d_B D(XY||XY_\star)} \quad (38)$$

where $\bar{X} \sim \bar{P}_B$, where $\bar{d}_B$ denotes the number of non-$\star$ symbols in $c(m)$. Note that the second equality in (38) holds uniformly over the set $\{\tilde{Q} : \tilde{Q} \approx Q\}$ by the continuity of divergence.[16]

Therefore,

$$\mathbb{E}(Z(m, \tilde{Q})) \doteq \frac{A}{d_B} 2^{-d_B D(XY||XY_\star)}. \quad (39)$$

Since $A = 2^{\beta B}$, from (26), (34), and (39) we get

$$\mathbb{E}_m(Z(m, \tilde{Q})) \doteq 2^{\eta B}.$$

Since $Z(m, \tilde{Q})$ is a binomial random variable, it can easily be seen from Chebyshev's inequality (or the Chernoff bound) that $Z(m, \tilde{Q})$ must be concentrated near its mean, from which it follows that

$$\mathbb{E}_m\left(\frac{Z(m, \tilde{Q})}{Z(m, \tilde{Q}) + 1}\right) = 1 - o(1) \quad B \to \infty. \quad (40)$$

From (37) and (40) we get

$$\begin{aligned}\mathbb{P}_m(\mathcal{E}) &\geq (1 - o(1)) \sum_{\{\tilde{Q}:\tilde{Q}\approx Q\}} \mathbb{P}_m(\bar{Y}_\nu^{\nu+\bar{d}_B-1} \in \mathcal{T}_{\tilde{Q}}(\bar{c}(m)))\\ &= 1 - o(1)\end{aligned} \quad (41)$$

as $B \to \infty$, where the second equality follows from Chebyshev's inequality. We conclude that for the single message case, the error probability tends to one whenever (34) holds.

2) Multiple messages: The main additional ingredient used to establish (29) is the fact that the decoder does not know a priori the transmitted message. Because of this, the augmented decoder's task is more difficult to perform; pure noise can induce an error whenever it generates a block that is typical with *any* of the (extended) codewords from $\mathcal{C}'_B$. The key element in the analysis consists in showing that the "typicality" regions associated with different codewords are essentially disjoint, *i.e.*, that the probability of the noise generating a block typical with any message is essentially $|\mathcal{C}'_B|$ times the probability for the single message case. This, together with the above argument for the single message case, yields the desired result.

Observe that since $\mathcal{C}'_B$ achieves a maximum error probability on the asynchronous channel that is less than ε_B, the (extended) codewords $\mathcal{C}'_B$ can also achieve a maximum error probability on the synchronous channel that is less than ε_B—if we reveal ν to the decoder, the channel becomes synchronous, and the error probability does not increase. Therefore, assuming that the decoder is deterministic, we can assign *disjoint* decoding regions $D(m)$ to each codeword of $\mathcal{C}'_B$ such that, with probability at least $1 - \varepsilon_B$, after transmission over the synchronous channel Q, the channel output lies in the decoding region $D(m)$ assigned to the transmitted codeword $c(m)$. If the decoder of $\mathcal{C}'_B$ is randomized, one can easily construct an expurgated code with a deterministic decoder and asymptotically the same rate as follows. Since the maximum error probability of $\mathcal{C}'_B$ is at most ε_B, the average error probability is at most ε_B, hence the average error probability under MAP decoding is also at most ε_B (note that MAP decoding minimizes the average error probability, not necessarily the maximum error probability). Now, without

[16] See foonote 13.

loss of optimality, the MAP decoder can be restricted to be deterministic. If we remove the half of the codewords with the largest error probability, we remain with a code whose maximum error probability is at most $2\varepsilon_B$ under a deterministic (MAP) decoding. This expurgated code and its decoding regions $\{D(m)\}$ can now be used for the argument.

Adapting the argument used for the single message case, fix a conditional distribution $\tilde{Q} \approx Q$ (see (36)), and let $Z(m,\tilde{Q})$ denote the binomial random variable representing the number of pure noise blocks that induce the conditional empirical distribution $\tilde{Q}$ with $\bar{c}(m)$. For each message m, define $D(m,\tilde{Q})$ as the intersection of the decoding region $D(m)$ with $\mathcal{T}_{\tilde{Q}}(\bar{c}(m))$—that is the set of sequences $y_1, y_2, \ldots, y_{d_B}$ in $D(m)$ whose y_i's corresponding to the non-$\star$ symbols of $c(m)$ have an empirical distribution $\tilde{Q}$ given $\bar{c}(m)$. Note that since the decoding regions are disjoint, the sets $D(m,\tilde{Q})$ are also disjoint.

Define

$$Z(\tilde{Q}) \triangleq \sum_m Z(m,\tilde{Q}),$$

and

$$D(\tilde{Q}) \triangleq \cup_m D(m,\tilde{Q})\,.$$

Then,

$$\begin{aligned}
\mathbb{E}[Z(\tilde{Q})] &= \sum_m \mathbb{E}[Z(m,\tilde{Q})] \\
&= (r_B-1)\sum_m \mathbb{P}_\star(D(m,\tilde{Q})) \\
&= (r_B-1)\sum_m 2^{-d_B(D(XY||XY_\star)+o(1))}\mathbb{P}_m(D(\tilde{Q},m)) \\
&= \frac{A}{d_B}2^{-d_B D(XY||XY_\star)(1+o(1))}\sum_m \mathbb{P}_m(D(m,\tilde{Q})) \\
&= \frac{A}{d_B}2^{-d_B D(XY||XY_\star)(1+o(1))}2^B\mathbb{P}(D(M,\tilde{Q})), \qquad (42)
\end{aligned}$$

where $\mathbb{P}_\star$ denotes the output distribution corresponding to $\bar{d}_B$ symbols $\star$; and where $\mathbb{P}_m$ denotes the output distribution when the channel input is $\bar{c}(m)$.

The first equality in (42) follows from the definition of $Z(\tilde{Q})$. The second equality follows from the definition of $Z(m,\tilde{Q})$ and the fact that there are r_B-1 pure noise blocks (see (32)). The third equality in (42) holds since the probability under $\mathbb{P}_\star$ of any sequence in $D(\tilde{Q},m)$ is equal to $2^{-d_B(D(XY||XY_\star)+o(1))}$ times the probability of that sequence under $\mathbb{P}_m$. To see this note that for any $y \in D(\tilde{Q},m)$ we have [4, Lemma 2.6]

$$\mathbb{P}_m(y) = 2^{-\bar{d}_B(H(\tilde{Y}|\bar{X})+D(\bar{X}\tilde{Y}||\bar{X}Y))}$$

and

$$\mathbb{P}_\star(y) = 2^{-\bar{d}_B(H(\tilde{Y})+D(\tilde{Y}||Y_\star))}\,.$$

Hence,

$$\begin{aligned}
\mathbb{P}_\star(y) &= \mathbb{P}_m(y)2^{-\bar{d}_B(D(\bar{X}\tilde{Y}||\bar{X}Y_\star)-D(\bar{X}\tilde{Y}||\bar{X}Y))} \\
&= \mathbb{P}_m(y)2^{-d_B(D(XY||XY_\star)+o(1))},
\end{aligned}$$

since $\tilde{Q} \approx Q$, by continuity of divergence.[17] The fourth equality in (42) follows from (33). For the fifth inequality in (42) we defined

$$\mathbb{P}(D(M,\tilde{Q})) \triangleq \frac{1}{|\mathcal{C}'_B|}\sum_m \mathbb{P}_m(D(m,\tilde{Q}))\,,$$

the average probability of successful decoding of the code $\mathcal{C}'_B$ and having an input/output joint type equal to $P_B\tilde{Q}$—in the above definition, M denotes the random message to be transmitted.

Now, recall that the probability of successful decoding of $\mathcal{C}'_B$ is at least $1-\varepsilon_B$ (see paragraph after (32)), hence

$$\mathbb{E}_{\hat{Q}}\mathbb{P}(D(M,\hat{Q})) \geq 1-\varepsilon_B\,.$$

Therefore, by Markov's inequality,

$$\mathbb{P}(\hat{Q} : \mathbb{P}\big(D(M,\hat{Q})\big) \geq 1-\sqrt{\varepsilon_B}\big) \geq 1-\sqrt{\varepsilon_B}\,,$$

i.e., with probability $1-\sqrt{\varepsilon_B} = 1-o(1)$, the empirical channel $\hat{Q}$ yields a probability of successful decoding $1-\sqrt{\varepsilon_B} = 1-o(1)$. Denoting by $\{\tilde{Q}\sim Q\}$ the set of conditional distributions $\tilde{Q}$ such that $\tilde{Q}\approx Q$ (see (36)) and such that

$$\mathbb{P}(D(M,\hat{Q})) \geq 1-\sqrt{\varepsilon_B}\,,$$

it follows that

$$\mathbb{P}(\hat{Q}\sim Q) = 1-o(1), \qquad (43)$$

since $\mathbb{P}(\hat{Q}\approx Q) = 1-o(1)$. Hence, from (42) and (33), we get

$$\mathbb{E}[Z(\tilde{Q})] = 2^{B(\beta-\delta)}2^{-d_B D(XY||XY_\star)(1+o(1))}2^B \qquad (44)$$

uniformly over $\{\tilde{Q}\sim Q\}$. Hence, if for some $\eta>0$ we have

$$B(1+\beta-\delta-\eta) > d_B D(XY||XY_\star)\,, \qquad (45)$$

then $\mathbb{E}Z(\tilde{Q}) \doteq 2^{\eta B}$, and using that $Z(\tilde{Q})$ is a binomial random variable, we get

$$\mathbb{E}\left(\frac{Z(\tilde{Q})}{Z(\tilde{Q})+1}\right) = 1-o(1)\,.$$

Proceeding as in (37), the error probability (averaged over messages) of the augmented decoder is lower bounded as

$$\begin{aligned}
\bar{\mathbb{P}}(\mathcal{E}) &\geq \sum_{\{\tilde{Q}:\tilde{Q}\sim Q\}} \bar{\mathbb{P}}(\mathcal{E}|\hat{Q}=\tilde{Q})\mathbb{P}(\bar{Y}(M)\in\mathcal{T}_{\tilde{Q}}(\bar{c}(M))) \\
&\geq \sum_{\{\tilde{Q}:\tilde{Q}\sim Q\}} \mathbb{E}\left(\frac{Z(\tilde{Q})}{Z(\tilde{Q})+1}\right)\mathbb{P}(\bar{Y}(M)\in\mathcal{T}_{\tilde{Q}}(\bar{c}(M))) \\
&= (1-o(1))\,. \qquad (46)
\end{aligned}$$

Hence, if (45) holds for some $\eta>0$, or, equivalently, if

$$\boldsymbol{R}\mathbb{E}(k(X))(1+\beta-\delta-\eta) > D(XY||XY_\star)$$

since $B/d_B = \boldsymbol{R}\mathbb{E}(k(X))$, the error probability tends to one as $B\to\infty$. This implies that if a code achieves rate $\boldsymbol{R}>0$ at timing uncertainty per information bit β and delay exponent

[17] See footnote 13.

δ then (29) holds. This completes the proof of the converses for Theorems 1 and 4. ∎

Converse of Theorem 5: The converse proof for Theorem 5 is almost the same as the the converse proof for Theorem 1. As for the achievability proofs, the main idea is to find a suitable replacement for the set $\{1, 2, \ldots, A\}$ of time slots that the receiver needs to consider. For the proof, we choose the set of time slots as a function of the coding scheme under consideration. In more detail, given any reliable coding scheme, *i.e.*, any coding scheme for which the probability of error $\varepsilon_B \to 0$ as $B \to \infty$, for each value t, consider the probability that the decoder makes an error or has delay greater than d_B conditioned on the event $\nu = t$. We will replace the set $\{1, \ldots, A\}$ with the set $\mathcal{S}(\sqrt{\varepsilon_B})$ of times t for which this conditional probability is at most $\sqrt{\varepsilon_B}$. Observe that the conditional probability of error, averaged over ν, is by definition at most $2\varepsilon_B$, so Markov's inequality says that the probability (over the distribution of ν) that this conditional probability is larger than $\sqrt{\varepsilon_B}$ is at most $2\sqrt{\varepsilon_B}$. Thus, ν is in $\mathcal{S}(\sqrt{\varepsilon_B})$ with probability at least $1 - 2\sqrt{\varepsilon_B}$. The key property of this construction is that the decoder for the given coding scheme can with high probability correctly decode the message within a delay of d_B for *each member* of $\mathcal{S}(\sqrt{\varepsilon_B})$.

We now apply the converse proof of Theorem 1 to the set $\mathcal{S}(\sqrt{\varepsilon_B})$. First, we need to parse the output sequence appropriately, *i.e.*, split the output sequence into disjoint blocks of length d_B. Recall that r_B, the number of such disjoint blocks, was roughly $\frac{A}{d_B}$ in the converse proof of Theorem 1. Now, however, since $\mathcal{S}(\sqrt{\varepsilon_B})$ can be arbitrary, it is possible that $\mathcal{S}(\sqrt{\varepsilon_B})$ does not even contain any time slots congruent to, say, 0 mod d_B. To get around this minor technicality, observe that by the pigeonhole principle, for at least one value x mod d_B, $\mathcal{S}(\sqrt{\varepsilon_B})$ contains at least $\frac{|\mathcal{S}(\sqrt{\varepsilon_B})|}{d_B}$ time slots congruent to x mod d_B. For such an x, we choose ν uniformly from those elements in $\mathcal{S}(\sqrt{\varepsilon_B})$ that are congruent to x mod d_B. Because the decoder for the given coding scheme can with high probability correctly decode the message within a delay of d_B for each member of $\mathcal{S}(\sqrt{\varepsilon_B})$, it follows that this decoder can decode the message and determine the value of ν with high probability even when ν is chosen as above.

From this point, we follow the converse proof of Theorem 1, with r_B replaced by $\frac{|\mathcal{S}(\sqrt{\varepsilon_B})|}{d_B}$ (equivalently, A is replaced by the size of $\mathcal{S}(\sqrt{\varepsilon_B})$). At the end, we see that a reliable decoder can exist only if for any $\eta > 0$ and B large enough,

$$B\left(1 + \frac{\log(\mathcal{S}(\sqrt{\varepsilon_B}))}{B}\right) \leq d_B(D(XY||XY_\star) + \eta).$$

Thus, $\frac{\log(\mathcal{S}(\sqrt{\varepsilon_B}))}{B}$ has replaced the role played by β in the converse proof of Theorem 1. Finally, since $\varepsilon_B \to 0$, $\sqrt{\varepsilon_B} \to 0$, so by definition of $\bar{\beta}$

$$\bar{\beta} \leq \limsup \frac{\log(\mathcal{S}(\sqrt{\varepsilon_B}))}{B},$$

completing the proof. ∎

Proof of Theorem 2: Starting from Theorem 1,

$$\boldsymbol{C}(\beta) = \max_X \min\left\{\frac{I(X;Y)}{\mathbb{E}[k(X)]}, \frac{I(X;Y) + D(Y||Y_\star)}{\mathbb{E}[k(X)](1+\beta)}\right\}. \quad (47)$$

A simple upper bound is

$$\begin{aligned}\boldsymbol{C}(\beta) &\leq \max_X \frac{I(X;Y) + D(Y||Y_\star)}{\mathbb{E}[k(X)](1+\beta)} &(48)\\ &= \frac{1}{1+\beta}\max_X \frac{\mathbb{E}[f(X)]}{\mathbb{E}[k(X)]}, &(49)\end{aligned}$$

where $f(x)$ is the divergence between the distribution of Y conditioned on $X = x$ and the distribution of Y conditioned on $X = \star$.

Using the fact that for nonnegative a, b, c, and d (with a suitable convention for the case where c and/or d is 0)

$$\frac{a+b}{c+d} \leq \max\left(\frac{a}{c}, \frac{b}{d}\right),$$

we see that the above maximum is achieved for an input distribution with a point mass at a^*, where

$$a^* = \text{argmax}_x \frac{f(x)}{k(x)}.$$

However, the maximizing solution is not unique. Since $f(\star) = k(\star) = 0$,

$$\frac{pf(\star) + (1-p)f(a^*)}{pk(\star) + (1-p)k(a^*)} = \frac{f(x)}{k(x)}$$

for any $p \in [0, 1]$. Hence, any input distribution with two point masses, one at $\star$ and one at a^*, will do. Going back to (49), we get

$$\boldsymbol{C}(\beta) \leq \frac{1}{1+\beta}\max_x \frac{f(x)}{k(x)}.$$

This upper bound is obtained by choosing the input distribution to maximize the second term in the minimum of (47). To prove that this upper bound can be achieved, choose X to have a distribution with probability p of being $\star$, and probability $1-p$ of being a^*, where $p \to 1$. The first term in the min approaches

$$\max_x \frac{f(x)}{k(x)}$$

by Theorem 3 of [8]. The second term is

$$\frac{1}{1+\beta}\max_x \frac{f(x)}{k(x)},$$

as derived above (true actually for any p, not only $p \to 1$). So, the second term is smaller, and we are always limited by the timing uncertainty. This proves the desired result. ∎

Remark. *Our results hold under the assumption that the only possible zero cost symbol is the $\star$ symbol. The other cases, which we now briefly discuss, can be handled with arguments similar to the ones used in this paper.*

- *Two symbols in $\mathcal{X}$ have zero cost: the capacity per unit cost is readily seen to be infinite.*
- *$\star \in \mathcal{X}$ and all $x \in \mathcal{X}$ have positive cost: the analysis in this paper can be applied, but would require some slightly cumbersome notation.*

- *There is a single zero cost symbol $x \in \mathcal{X}$ different than $\star$: in this case the asynchronous capacity per unit cost is*

$$\boldsymbol{C}(\beta,\delta) = \frac{\boldsymbol{C}(0,0)}{1-2\delta} \quad 0 \le \delta \le \beta/2,\ \beta \ge 0\,,$$

i.e., it is the synchronous capacity per unit cost multiplied by a factor $1/(1-2\delta)$.
The first thing to note in the above capacity expression is that it does not depend on β. The reason for this is that no matter how large β is it is always possible to append to each codeword a long enough zero cost preamble that guarantees the decoder is able to identify σ with high probability.
For an intuitive justification for the $1-2\delta$ factor, observe that in the achievability proof of Theorem 4 case b., δB bits are encoded via σ, the start *information time. When a symbol different than $\star$ has zero cost, not only it is possible to encode information through the start information time, but also in the codeword "length." By codeword length we mean the time between σ and the time of the last non-zero cost symbol of the sent codeword. This allows to communicate $2\delta B$ of information only through timing.*

Proof of Theorem 3: A simple quantization argument can be used to derive Theorem 3 from Theorem 1. For achievability, one quantizes the input and the output real values to a finite alphabet. Then, the achievability part of Theorem 1 can be applied to this quantized channel. Finally, take the limit of infinitely fine quantization to proves that the stated rate is achievable.

For the converse, one adapts the method of types by quantizing the set of probability distributions, *i.e.*, one defines a type as a set of probability distributions that are "close" to each other. With such a notion of type, the converse part of Theorem 1 can be applied, and in the limit of infinitely fine quantization, one obtains the desired converse result. ■

REFERENCES

[1] V. Chandar, A. Tchamkerten, and G.W. Wornell, *Optimal sequential frame synchronization*, IEEE Trans. Inform. Th. **54** (2008), no. 8, 3725–3728.

[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] T.M. Cover and J.A. Thomas, *Elements of information theory*, Wiley, New York, 2006.

[4] I. Csiszàr and J. Körner, *Information theory: Coding theorems for discrete memoryless channels*, Cambridge University Press, New York, 2011.

[5] R. G. Gallager, *Energy limited channels: Coding, multiaccess, and spread spectrum*, report, M.I.T., Laboratory for Information and Decision Systems, 1987.

[6] A. Tchamkerten, V. Chandar, and G. Wornell, *Asynchronous communication: capacity bounds and suboptimality of training*, submitted to IEEE Trans. Inform. Th. (2011).

[7] A. Tchamkerten, V. Chandar, and G.W. Wornell, *Communication under strong asynchronism*, IEEE Trans. Inform. Th. **55** (2009), no. 10, 4508–4528.

[8] S. Verdú, *On capacity per unit cost*, IEEE Trans. Inform. Th. **36** (1990), no. 5, 1019–1030.

Venkat Chandar received S.B. degrees in EECS and mathematics in 2006, an M. Eng. in EECS in 2006, and a Ph. D. in EECS in 2010, all from MIT. His current research interests include coding theory and algorithms, with an emphasis on the construction and analysis of sparse graph codes for various problems related to communication, compression, sensing, and information-theoretic secrecy.

Aslan Tchamkerten received the Engineer Physicist Diploma in 2000 and the Ph.D. degree in Communications in 2005, both from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Between 2005 and 2008, he was a Postdoctoral Associate in the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge. In 2008 he joined Telecom ParisTech (ex. Ecole Nationale Supérieure des Télécommunications, ENST), Paris, France, where he is currently Associate Professor. In 2009, he won a junior excellence chair grant from the French National Research Agency (ANR). His research interests are in information theory, applied Statistics, and algorithms.

David Tse received the B.A.Sc. degree in systems design engineering from University of Waterloo in 1989, and the M.S. and Ph.D. degrees in electrical engineering from Massachusetts Institute of Technology in 1991 and 1994 respectively. From 1994 to 1995, he was a postdoctoral member of technical staff at A.T. & T. Bell Laboratories. Since 1995, he has been at the Department of Electrical Engineering and Computer Sciences in the University of California at Berkeley, where he is currently a Professor. He received a 1967 NSERC graduate fellowship from the government of Canada in 1989, a NSF CAREER award in 1998, the Best Paper Awards at the Infocom 1998 and Infocom 2001 conferences, the Erlang Prize in 2000 from the INFORMS Applied Probability Society, the IEEE Communications and Information Theory Society Joint Paper Award in 2001, the Information Theory Society Paper Award in 2003, the 2009 Frederick Emmons Terman Award from the American Society for Engineering Education, and a Gilbreth Lectureship from the National Academy of Engineering in 2012. He has given plenary talks at international conferences such as ICASSP in 2006, MobiCom in 2007, CISS in 2008, and ISIT in 2009. He was the Technical Program co-chair of the International Symposium on Information Theory in 2004, and was an Associate Editor of the IEEE Transactions on Information Theory from 2001 to 2003. He is a coauthor, with Pramod Viswanath, of the text "Fundamentals of Wireless Communication", which has been used in over 60 institutions around the world.