Asynchronous Capacity per Unit Cost

Venkat Chandar\textsuperscript{1}, Aslan Tchamkerten\textsuperscript{2}, and David Tse\textsuperscript{3}
\textsuperscript{1}Massachusetts Institute of Technology
\textsuperscript{2}Telecom ParisTech
\textsuperscript{3}University of California, Berkeley

Abstract—The capacity per unit cost, or equivalently minimum cost to transmit one bit, is a well-studied quantity. It has been studied under the assumption of full synchrony between the transmitter and the receiver. In many applications, such as sensor networks, transmissions are very bursty, with small 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 \textit{a priori} synchronization. In this paper, we show that the minimum cost to transmit $B$ bits of information asynchronously is $(B + H)k_{\text{sync}}$, where $k_{\text{sync}}$ is the synchronous minimum cost per bit and $H$ is a measure of timing uncertainty equal to the entropy for most reasonable arrival time distributions.

I. INTRODUCTION

Synchronization is an important component of any communication system. To understand the cost of synchronization on communication performance, 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 symbol synchronization, say by sending a pilot sequence, is relatively small in such applications because the cost can be amortized over the many symbols transmitted. In the second type, transmissions are very bursty, with very small 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 \textit{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 \textsuperscript{4}, \textsuperscript{5}. 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 many 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 \textit{capacity per unit cost} \textsuperscript{3}, \textsuperscript{6}, 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 \textsuperscript{4}, \textsuperscript{5} 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 $R$ 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 $\nu$, not known \textit{a priori} to the receiver. Without knowing $\nu$, 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 $\nu$, we assume that both the transmitter and the receiver know that $\nu$ 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 the 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: for $B$ large, the minimum cost to transmit $B$ bits of information asynchronously is:

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

where $k_{\text{sync}}$ is the minimum cost to transmit one bit of information in the synchronous setting. \textsuperscript{4} 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 $A$ is at least exponential $B$.

\textsuperscript{1}In this paper, all logarithms are taken to base 2.
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 be no price to pay for the timing uncertainty, but the delay incurred would be very large, exponential in $B$. In contrast, the above result can be achieved by a coding scheme whose delay is much shorter, linear in the number of bits $B$, and we show that performance cannot be improved within the broad class of coding schemes whose delays are sub-exponential in $B$. A delay linear in $B$ is of the same order as the delay incurred in the synchronous mission is highly random to the receiver and the additional cost when the allowable delay is no larger than $A$, the minimum cost to transmit $B$ bits can be further reduced to:

$$B + \log \left( \frac{A}{D} \right)$$

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

**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 $\mathcal{X} \cup \{\star\}$ and $\mathcal{Y}$, respectively, and transition probability matrix $Q(y|x)$ for all $x \in \mathcal{X} \cup \{\star\}$ and $y \in \mathcal{Y}$. 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 $\star$.

Given $B$ information bits to be transmitted, a codebook $c$ consists of $M = 2^B$ codewords of length $N$ composed of symbols from $\mathcal{X}$. The message $m$ arrives at the transmitter at a random time $\nu$, 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 $\sigma(\nu, m)$ to begin transmitting the codeword $c(m) \in c$ assigned to message $m$—the transmitter need not start sending information right at the time when the message is available, i.e., at time $\nu$. The only constraint $\sigma$ must satisfy is that

$$\nu \leq \sigma(\nu, m) \leq A,$$

i.e., the transmitter cannot start transmitting before the message arrives or after the end of the uncertainty window. 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 $\{\nu = t\}$, $t \in \{1, 2, \ldots, A\}$, 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 $\sigma(t, m) + N \leq i \leq \sigma + A + N - 1$, 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 $i$th symbol of the codeword $c(m)$.

Knowing the asynchronism level $A$, but not the value of $\nu$, the receiver decodes by means of a sequential test $(\tau, \phi)$, where $\tau$ 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 $\phi$ denotes a decision rule that declares the decoded message (see Fig. 1). 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 $t \geq 1$. Given $\{\tau = t\}$, $t \in \{1, 2, \ldots, A+N-1\}$, the function $\phi$ outputs a message based on the past observations from time 1 up to time $t$.

A ‘code’ refers to a codebook $c$ together with a decoder, i.e., a sequential test $(\tau, \phi)$. Throughout the paper, whenever clear from context, we often refer to a code using the codebook symbol $c$ only, letting the decoder implicit.

The maximum (over messages) decoding error probability for a given code $c$ is defined as

$$P(E|c) = \max_m \frac{1}{A} \sum_{i=1}^{A} P_{m,i}(c),$$

where $P_{m,i}(c)$ is the $i$th component of the probability distribution of message $m$ under the code $c$.
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.

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

**Definition 2** (Cost of a Code). The (maximum) cost of a code \(\mathcal{C}\) is defined as
\[
K(\mathcal{C}) \triangleq \max_m \sum_{i=1}^{N} k(c_i(m)).
\]

**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
\[
\min_m \mathbb{P}_m(\tau - \nu \leq d) \geq 1 - \varepsilon,
\]
where \(\mathbb{P}_m\) denotes the output distribution conditioned on the sending of message \(m\).

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 \(B \rightarrow \infty\) while \(\beta\) is kept fixed.

**Definition 4** (Asynchronous Capacity Per Unit Cost). \(R\) is an achievable rate per unit cost at timing uncertainty per information bit \(\beta\) and delay exponent \(\delta\) if there exists a sequence of codes \(\{\mathcal{C}_B\}\), and a sequence of numbers \(\{\varepsilon_B\}\) with \(\varepsilon_B \rightarrow \infty\) 0, such that
\[
\mathbb{P}(\mathcal{C}_B) \leq \varepsilon_B, \quad \limsup_{B \rightarrow \infty} \frac{\log(D(\mathcal{C}_B, \varepsilon_B))}{B} \leq \delta,
\]
and
\[
\liminf_{B \rightarrow \infty} \frac{B}{K(\mathcal{C}_B)} \geq R.
\]
The asynchronous capacity per unit cost, denoted by \(C(\beta, \delta)\), is the largest achievable rate per unit cost. In the important case when \(\delta = 0\), we define \(C(\beta) \triangleq C(\beta, 0)\).

Note that, in Definition 3, the codeword length \(N\) is a free parameter that can be optimized, just as for the synchronous capacity per unit cost (see comment after 2 Definition 2). The results in the next section characterize the capacity per unit cost for arbitrary \(\beta\) and \(\delta\) and arbitrary alphabet \(\mathcal{X}\). Similarly as for the synchronous case, the results simplify when there is a zero cost symbol, specifically when \(\mathcal{X}\) contains \(\star\) and \(\star\) has zero cost.

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 [6], which states that the synchronous capacity per unit cost is
\[
\max_{\mathcal{X}} \frac{I(X; Y)}{\mathbb{E}[k(X)]}.
\]

**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
\[
C(\beta) = \max_{\mathcal{X}} \min_{\varepsilon} \left\{ \frac{I(X; Y)}{\mathbb{E}[k(X)]}, \frac{I(X; Y) + D(Y||Y_*)}{\mathbb{E}[k(X)](1 + \beta)} \right\},
\]
where \(X\) denotes the random input to the channel, \(Y\) the corresponding output, \(Y_*\) 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 \(D(Y||Y_*)\) the Kullback-Leibler distance between the distributions of \(Y\) and \(Y_*\).

Furthermore, capacity can be achieved by codes whose delay grows linearly with \(B\).

The two terms in (3) 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. To see this, note that by swapping the max and the min in (3), we deduce that \(C(\beta)\) is less than (2), the synchronous capacity per unit cost.

The second term in (3) corresponds to the receiver’s ability to determine the arrival time \(\nu\) of the data. Indeed, even though the decoder is required only 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 (approximately) locate the time codeword transmission begins—the delay constraint imposes the decoder to locate the sent message within a time window that is negligible compared to \(A\). The quantity \(I(X; Y) + D(Y||Y_*)\) 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 (2) simplifies, and Theorem 3 in (6) says that the synchronous capacity per unit cost becomes
\[
\max_{x \in \mathcal{X}} \frac{D(Y_x||Y_0)}{k(x)},
\]
an optimization over the input alphabet instead of over the set of all input distributions, where \(Y_x\) refers to is 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:

3Throughout the paper we use the standard ‘big O’notation to characterize growth rates.

4\(Y_*\) can be interpreted as ‘pure noise.’
**Theorem 2** (Asynchronous Capacity per Unit Cost With Zero Cost Symbol: Sub-exponential Delay Constraint). If \( \ast \) is in \( \mathcal{X} \) and has zero cost, the asynchronous capacity per unit cost at delay exponent \( \delta = 0 \) is given by

\[
C(\beta) = \frac{1}{1 + \beta} \max_{x \in \mathcal{X}} \frac{D(Y_x||Y_\ast)}{k(x)},
\]

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. This allows us to consider message/location pairs as inducing a code of size \( \frac{2}{\beta} B^2 \) used for communication across the synchronous channel. Hence, if, say, \( N \) grows sub-exponentially with \( B \), 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 \( \beta \) 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 PPM. From [6], 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 \( (1 + \beta) k_{sym} \) is an achievable rate per unit cost.

Theorem 2 can be extended to the (continuous-valued) Gaussian channel, where the idle symbol \( \ast \) 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

\[
C(\beta) = \frac{1}{1 + \beta} \frac{\log e}{N_0} \beta \geq 0.
\]

Theorem 1 can be extended to the case of a large delay constraint, i.e., when \( 0 < \delta < \beta \). As for Theorem 1 the following result holds irrespectively of whether or not \( \mathcal{X} \) contains \( \ast \). A simplification similar to Theorem 2 applies if \( \mathcal{X} \) contains \( \ast \) and it has zero cost.

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

\[
C(\beta, \delta) = C(\beta - \delta),
\]
i.e., it is the same as the capacity per unit cost with delay exponent \( \delta = 0 \), but with asynchronism exponent \( \beta \) reduced to \( \beta - \delta \).

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

Below, \( \nu^B \) denotes the arrival time random variable when \( B \) bits of information have to be transmitted (in Theorem 1 \( \nu^B \) has the uniform distribution over \( \{1, 2, \ldots, 2^B\} \).

**Theorem 5** (Asynchronous Capacity per Unit Cost With Non-uniform Arrival Time: Sub-exponential Delay Constraint). Define

\[
\overline{\beta} = \inf_{\epsilon_B} \lim_{B \to \infty} \frac{\log(S(\epsilon_B))}{B},
\]

where the infimum is with respect to all sequences \( \{\epsilon_B\} \) of nonnegative numbers such that \( \lim_{B \to \infty} \epsilon_B = 0 \), where \( S(\epsilon_B) \) denotes the size of the smallest set with probability at least \( 1 - \epsilon_B \), and it is assumed that the limit in (6) exists.

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

\[
C(\overline{\beta}) \geq \max_{X} \min \left\{ \frac{I(X;Y)}{E[k(x)]} \bigg/ \frac{I(X;Y) + D(Y||Y_\ast)}{E(k(x))(1 + \beta)} \right\}.
\]

Although the formula for \( \overline{\beta} \) in (6) appears unwieldy, in many cases it can easily be evaluated. For example, in many cases, such as for the uniform or the geometric distributions, the formula reduces to the normalized entropy

\[
\overline{\beta} = \lim_{B \to \infty} H(\nu^B)/B.
\]

There are cases, however, where (6) 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 \( (1/2)(1 - \beta)^{B-i} \) for \( i = 2, \ldots, 2^B + 1 \). Then,

\[
\overline{\beta} = 2 \lim_{B \to \infty} H(\nu^B)/B.
\]

**IV. PROOFS OF RESULTS**

**Achievability of Theorem 7**. We first show the existence of a random code with the desired properties. Then, via an expurgation argument, we show the existence of a deterministic code achieving the same (asymptotic) performance as the random code.

Fix some arbitrary distribution \( P \) on \( \mathcal{X} \). Let \( X \) be the input having that distribution and let \( Y \) be the corresponding output. Given \( B \) bits of information to be transmitted, the codebook \( \mathcal{E} \) 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 \( x^N \) satisfies the following ‘constant composition’ property

\[
||P_{x^N} - P|| \leq 1/\log N,
\]

we let \( c(m) = x^N \). Otherwise, we repeat the procedure until we generate a sequence sufficiently close to \( P \). It is not hard to see that, for a fixed \( m \), no repetition will be required to generate \( c(m) \): the constant composition property holds with probability tending to one as \( N \to \infty \). The obtained codebook is thus essentially of constant composition, i.e., each symbol appears roughly the same number of times across codewords.
The sequential typicality decoder operates as follows. At time $t$, for all $m \in \{1, 2, \ldots, 2^B\}$, the typicality decoder computes the empirical distributions

$$P_{C(m), Y_{t-N+1}}(\cdot, \cdot)$$

induced by $C(m)$ and the $N$ output symbols $Y_{t-N+1}$. If there is a unique message $m$ for which

$$||P_{C(m), Y_{t-N+1}}(\cdot, \cdot) - P(\cdot)Q(\cdot)|| \leq 1/\log N,$$

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 at random.

Induced by the union of two error events:

- $E_1$: the decoder stops at time $t$ between $\nu$ and $\nu + 2N - 2$ (including $\nu$ and $\nu + 2N - 2$), and declares $m'$;
- $E_2$: the decoder stops either at some time $t$ before $\nu$ or from $\nu + 2N - 1$ onwards, and declares $m'$.

For the first error event, for some $0 \leq k \leq N - 1$ 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. The probability that such a $Y^N$ yields an type $J$ that is jointly typical with $P(\cdot)Q(\cdot)$, that is

$$||J(\cdot, \cdot) - P(\cdot)Q(\cdot)|| \leq 1/\log N,$$

is upper bounded as

$$P_m(P_{C(N), Y_N} = J) = \sum_{y^N \in Y^N} P(Y^N = y^N) \sum_{x^N : P_{x, y^N} = J} P(X^N = x^N) \leq \sum_{y^N \in Y^N} P(Y^N = y^N) \sum_{x^N : P_{x, y^N} = J} 2^{-N(H(J_X) + D(J_X || P) - \varepsilon)} \leq \sum_{y^N \in Y^N} P(Y^N = y^N) 2^{-(NH(J_X) - \varepsilon)} \{x^N : P_{x, y^N} = J\} \leq \sum_{y^N \in Y^N} P(Y^N = y^N) 2^{-(NH(J_X) - \varepsilon)} 2^{NH(J_X | y)} \leq 2^{-N(I(J) - \varepsilon)} \leq 2^{-N(I(X; Y) - 2\varepsilon)}$$

for any $\varepsilon > 0$ and all $N$ large enough, where $H(J_X)$ denotes the entropy of the left marginal of $J$.

We use here capital letters to denote codewords to emphasize that they are randomly generated. $||\cdot||$ refers to the $L_1$-norm.

The notion of typicality we use is often referred to as ‘strong typicality’ in the literature.

Recall that the type of a string $y^N \in Y^N$, denoted by $P_{y^N}$, assigns to each $a \in \mathcal{Y}$ a probability that corresponds to the frequency of occurrences of $a$ within $y^N$ [2, Chapter 1.2]. For instance, if $y^2 = (0, 1, 0)$, then $P_{y^2}(0) = 2/3$ and $P_{y^2}(1) = 1/3$. The joint type induced by a pair of strings $x^N \in \mathcal{X}^N, y^N \in \mathcal{Y}^N$ is defined similarly.

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

The first equality in (7) 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 would get from [11, Theorem 11.1.2, p. 349]

$$P_{l.i.d.}(X^N = x^N) = 2^{-N(H(J_X) + D(J_X || P))}.$$ 

Since a codeword with each component generated i.i.d. according to $P$ satisfies the constant composition property with probability tending to one as $N \to \infty$, we get

$$P(X^N = x^N) = 2^{-N(H(J_X) + D(J_X || P)}(1 + o(1))$$ as $N \to \infty$, which justifies the second inequality. The third inequality follows from [2, Lemma 2.5, p. 31]. The fifth inequality holds for any $\varepsilon > 0$ and all $N$ large enough, since by assumption $J$ is close to $PQ$.

Thus, taking a union bound over time $(2N)$, we obtain the upper bound

$$P_m(E_1) \leq 2N \cdot 2^{-N(I(X; Y) - \varepsilon)},$$ valid for $N$ large enough.

For the second error event, pure noise produces some output $Y^N$ that is jointly typical with $C(m')$. From Sanov’s theorem and the continuity of the Kullback-Leibler distance (needed only in the first argument) the probability of this event happening at a particular time is upper bounded by

$$2^{-N(D(XY || XY) - \varepsilon)} = 2^{-N(I(X; Y) + D(\bar{P}_Y || P_Y) - \varepsilon)}$$

for all $N$ large enough. Here $D(XY || XY)$ refers to the Kullback-Leibler distance between, on the one hand, the joint distribution of $X$ and $Y$ and, on the other hand, the product of the distributions of $X$ and $Y$. Hence, taking a union bound over all times where noise could produce such output, we get

$$P_m(E_2) \leq A \cdot 2^{-N(I(X; Y) + D(\bar{P}_Y || P_Y) - \varepsilon)}$$

for all $N$ large enough.

Combining, we get

$$P_m(m \to m') \leq 2N \cdot 2^{-N(I(X; Y) - \varepsilon)} + A \cdot 2^{-N(I(X; Y) + D(\bar{P}_Y || P_Y) - \varepsilon)}.$$ Hence, taking a union bound over all possible wrong messages, we obtain for all $\varepsilon > 0$

$$P_m(E) \leq 2^B \left(2N \cdot 2^{-N(I(X; Y) - \varepsilon)} + A \cdot 2^{-N(I(X; Y) + D(\bar{P}_Y || P_Y) - \varepsilon)}\right)$$

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

$$\mathbb{E}_c(\mathbb{P}(E)) = P_m(E) \leq 2^B \left(2N \cdot 2^{-N(I(X; Y) - \varepsilon)} + A \cdot 2^{-N(I(X; Y) + D(\bar{P}_Y || P_Y) - \varepsilon)}\right) \triangleq \varepsilon_1(N).$$
where \( \bar{\mathbb{P}} \) denotes probability averaged over messages. We now show that the delay of our coding scheme in the sense of Definition 3 is \( N \). If \( \tau \geq \nu + N \) then necessarily it means that \( Y_{\nu+N-1}^{\nu} \) isn’t typical with the sent codeword. From Sanov’s Theorem, the probability of the latter event tends to zero as \( N \to \infty \). Therefore, denoting by \( D(\mathbb{C}, \varepsilon_2(N)) \) the delay averaged over messages\(^{11}\), we get

\[
\mathbb{P}(\bar{D}(\mathbb{C}, \varepsilon_2(N)) \leq N) = 1 - \varepsilon_2(N), \tag{9}
\]

where \( \varepsilon_2(N) \) is a function that tends to zero as \( N \to \infty \).

So far we have proved that a random code has error probability, averaged over messages, less than \( \varepsilon_1(N) \) and delay, averaged over messages, less than \( N \) with probability at least \( 1 - \varepsilon_2(N) \). Moreover, all codewords in our random ensemble have cost \( N \mathbb{E}[k(X)](1 + o(1)) \) as \( N \to \infty \).

We now show the existence of a non-random code with maximum delay close to \( N \) and maximum delay and error probability within a constant factor of \( \varepsilon_1(N) \) via a two-step expurgation procedure over codes and messages.

Define the events

\[
\mathcal{A}_1 = \{ \bar{\mathbb{P}}(\mathcal{E} | \mathbb{C}) \leq \varepsilon_1(N)(1 + \eta_1) \}
\]

\[
\mathcal{A}_2 = \{ D(\mathbb{C}, \varepsilon_2(N)) \leq N \}
\]

where \( \eta_1 > 0 \) is some arbitrary constant, and where \( \bar{\mathbb{P}}(\mathcal{E} | \mathbb{C}) \) denotes the average, over messages, error probability given code \( \mathbb{C} \) is used. We then deduce that

\[
\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) \geq 1 - \frac{1}{1 + \eta_1} - \varepsilon_2(N). \tag{10}
\]

We used Markov’s inequality to establish that

\[
\mathbb{P}(\mathcal{A}_1^\prime) \leq \frac{1}{1 + \eta_1}.
\]

From (10), for all \( N \) large enough there exists a non-random code \( \mathbb{C} \) whose average, over messages, delay and error probability satisfies

\[
D(\mathbb{C}, \varepsilon_2(N)) \leq N,
\]

and

\[
\bar{\mathbb{P}}(\mathcal{E} | \mathbb{C}) \leq \varepsilon_1(N)(1 + \eta_1).
\]

respectively. We now strengthen the above conclusion for maximum delay and error probability.

For an arbitrary small \( \eta_2 > 0 \), remove all the codewords in \( \mathbb{C} \) with delay larger or equal than \( (1 + \eta_2)N \). From the remaining set of codewords, at least

\[
2^{\beta B}(1 - 1/(1 + \eta_2))
\]

of them, remove the half of the codewords with the highest error probability. This final set \( \mathbb{C}' \) contains at least \( 2^{\beta B}(1 - 1/(1 + \eta_2)) \) codewords and has maximum error probability and maximum delay upper bounded as

\[
\mathbb{P}(\mathcal{E} | \mathbb{C}') \leq \varepsilon_1(N) \frac{2(1 + \eta_1)}{1 - 1/(1 + \eta_2)}, \tag{11}
\]

and

\[
D(\mathbb{C}', \varepsilon_2(N)) \leq N(1 + \eta_2), \tag{12}
\]

respectively.

Finally, recall that by construction, all the codewords have cost \( N \mathbb{E}[k(X)](1 + o(1)) \) as \( N \to \infty \). Hence, for \( N \) large enough

\[
k(\mathbb{C}') \leq N \mathbb{E}[k(X)](1 + \eta_2). \tag{13}
\]

Therefore, from (11), (12), and (13), our non-random code \( \mathbb{C}' \) has an error probability that tends to zero as \( \varepsilon_1(N) \) vanishes, a (maximum) delay upper bounded by \( N(1 + \eta_2) \), and a cost at most \( N \mathbb{E}[k(X)](1 + \eta_2) \).

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 (8)). Then, \( \mathbb{P}(\mathcal{E} | \mathbb{C}') \) goes to zero as \( B \to \infty \) provided that

\[
\frac{B}{N} < \min \left\{ \frac{I(X; Y)}{1 + \beta}, \frac{I(X; Y) + D(Y | Y_\sigma)}{1 + \beta} \right\}.
\]

Since the cost of \( \mathbb{C}' \) is at most \( N \mathbb{E}[k(X)](1 + \eta_2) \), the above condition is implied by the following condition

\[
\frac{B}{K(\mathbb{C}')} < \min \left\{ \frac{I(X; Y)}{(1 + \eta_2)\mathbb{E}[k(X)]}, \frac{I(X; Y) + D(Y | Y_\sigma)}{\mathbb{E}[k(X)](1 + \eta_2)(1 + \beta)} \right\}.
\]

Maximizing over all input distributions and using the fact that \( \eta_1 > 0 \) and \( \eta_2 > 0 \) are arbitrary proves that (3) is asymptotically achieved by non-random codes with delay linear in \( B \).

From the above analysis it is easy to see that whenever there exists some input \( X \) such that \( I(X; Y) > 0 \) while \( \mathbb{E}[k(X)] = 0 \), and thus \( X \) contains more than one zero cost symbol, the asynchronous capacity per unit cost is infinite. ■

Achievability of Theorem 4 The achievability scheme for Theorem 4 is similar to the achievability scheme used to prove Theorem 1. The only difference is that now the transmitter does not start transmitting at time \( \nu \). Instead, the transmitter first reduces the receivers’ time uncertainty about the beginning of codeword transmission by waiting to transmit until the first multiple of \( 2^B \) larger than \( \nu \). This effectively reduces the uncertainty about codeword transmission from \( 2^\beta B \) to \( 2^{(\beta - \delta)B} \), and one proves that \( C(\beta - \delta) \) is achievable with delay \( O(2^\beta B) \) by repeating the arguments of the achievability of Theorem 1 by also fixing the ratio \( B/N \). Hence, the blocklength is exponentially smaller than the delay. This is in contrast with the achievability of Theorem 1 where delay and blocklength are the same (information transmission starts at the same time that information arrives, i.e., \( \sigma(\nu, n) = \nu \)). ■

Achievability of Theorem 5 To prove the achievability part of Theorem 5, one applies the same arguments as for the achievability of Theorem 1 by replacing the uncertainty set \( A = \{1, 2, \ldots, A\} \) by a ‘typical’ set of size \( S(\varepsilon_B) \) whose probability, under the arrival time distribution, is at least \( 1 - \varepsilon_B \) (such as set exists by assumption). Note that the event when the arrival time \( \nu \) doesn’t belong to the typical set affects neither the cost nor the delay (asymptotically). ■

Converse of Theorem 1 Recall that delay refers to the elapsed time between \( \nu \) and \( \tau \), and need not coincide with the
codeword length $N$. Throughout this proof it is convenient to refer to `codeword’ the sequence of symbols transmitted from time $\nu$ until time $\nu + d - 1$, where $d$ is the achieved (maximum) delay. (Hence, with this definition a codeword is mostly composed of * if $N \ll d$.) Accordingly, a codebook represents now a collection of the newly defined codewords.

Assume that $\{C_B\}$ achieves a rate per unit cost $R > 0$ at timing uncertainty per unit cost $\beta$ and delay exponent $\delta = 0$. To simplify notation, we denote the delay $D(C_B, \varepsilon_B)$, where $\varepsilon_B \overset{B \to \infty}{\to} 0$, by $d_B$. By assumption (we consider Theorem 1), the delay exponent is zero, i.e.,

$$\limsup_{B \to \infty} \frac{\log d_B}{B} = 0.$$  

We now show that for any $\eta > 0$, $R$ and $\beta$ satisfy

$$RE[k(X)] \leq I(X,Y)(1 + \eta)$$  

and

$$RE[k(X)](1 + \beta) \leq D(XY|X,Y_*) + \eta$$  

for $B$ large enough, where $X \sim P_B$, and where $P_B$ denotes the type containing the most codewords from $C_B$.

First we prove (14). Let $C_B'$ be the constant composition subset of $C_B$ with the largest number of elements, and let $P_B$ be corresponding type. $C_B'$ is clearly a good code for the synchronous channel, i.e., if we reveal $\nu$ to the receiver and decoding happens at time $\nu + d_B - 1$, it is possible to achieve error probability less than $\varepsilon_B$. It follows from [2] Lemma 1.4, p. 104] that for any $\eta > 0$,  

$$\log |C_B'| \leq I(X;Y)(1 + \eta)$$  

for all $B$ large enough. Then, to obtain (14), we use the fact that the number of types grows polynomially with $d_B$ [2 Lemma 2.2, p. 28], and that the cost of $C_B'$ is equal to $d_B \cdot E[k(X)]$.

To prove (15), let us reveal the complete output sequence

$$y_1, y_2, \ldots, y_{A+N-1}$$  

to the receiver, and that the message was sent in one of the $r = \left\lfloor \frac{A + d_B - (\nu \mod d_B)}{d_B} \right\rfloor$ consecutive (disjoint) blocks of duration $d_B$ as shown in Fig. 2.

With this additional knowledge, the optimal MAP decoder is able to simultaneously output the sent codeword and the block of size $d_B$ corresponding to the actual transmission period, with probability at least $1 - 2^{-E_B} > 0$. To see this, note that $C_B$ achieves error probability $E_B$, and that the (maximum) communication delay is less than $d_B$ with probability at least $1 - \varepsilon_B$.

To develop some intuition for (15), first consider the case where the sent codeword is known at the decoder, so that the decoder’s task is only to output the block of size $d_B$ that corresponds to the period when the codeword is sent. We show that if $\beta$ is sufficiently large, the decoder will not be able to perform this 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 MAP decoder has a large probability of error whenever for some $\eta > 0$ and all $B$ large enough,

$$B\beta > d_B(D(XY|X,Y_*) + \eta).$$  

First note that the MAP decoder will fail with probability at least $1/2$ whenever a pure noise block of $d_B$ output symbols induces the same joint type with $c(m)$ as the block of output symbols $Y_{\nu}, \ldots, Y_{\nu+d_B-1}$. Since $c(m)$ has type $P_B$, the joint type of $c(m)$ and $Y_{\nu}, \ldots, Y_{\nu+d_B-1}$ is close to $P_B Q$ with probability $1 - o(1)$, i.e., $Y_{\nu}, \ldots, Y_{\nu+d_B-1}$ is (strongly) typical with $c(m)$ with high probability. Therefore, to show that the MAP decoder fails with probability bounded away from 0, it suffices to show that for any conditional type $\mathcal{Q} \approx \mathcal{Q}$, with high probability there exists at least one pure noise block that induces the joint type $P_B \mathcal{Q}$ with $c(m)$. Here $\mathcal{Q} \approx \mathcal{Q}$ means that

$$\|P_B \mathcal{Q} - P_B \mathcal{Q}\| \leq 1/\log d_B.$$  

From standard results in large deviations, the probability that one single pure noise block induces the joint type $P_B \mathcal{Q}$ with $c(m)$ is $\leq 2^{-d_B D(XY|X,Y_*)}$, where $\mathcal{Y}$ denotes the channel output when $X$ is the input to $Q$. Therefore, the number of pure noise blocks that induce the joint type $P_B \mathcal{Q}$ with $c(m)$ is a binomial random variable with mean

$$\frac{A}{d_B} 2^{-d_B D(XY|X,Y_*)} \leq \frac{A}{d_B} 2^{-d_B D(XY|X,Y_*)},$$

where for the second equality we used that $\mathcal{Q} \approx \mathcal{Q}$. Hence, since $d_B$ is sub-exponential in $B$, that $A = 2^{\beta B}$, from (17) it follows that the mean grows at least as $2^{\beta B}$, for some $\delta > 0$. Since the variance of a binomial random variable is at most its mean, Chebyshev’s inequality implies that with probability $1 - o(1)$ there is at least one pure noise block that induces the joint type $P_B \mathcal{Q}$ with $c(m)$. Therefore, the MAP decoder fails with probability at least $1/2 - o(1)$ whenever (17) holds. Hence, whenever there is only a single message and the decoder’s only task is to locate it, necessarily we have

$$B\beta \leq d_B(D(XY|X,Y_*) + \eta)$$  

for any $\eta > 0$ and all $B$ large enough.

We now extend the above argument to obtain inequality (15). To obtain (15) we use the fact that the decoder does not know a priori the transmitted message. Because of this,
the 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 messages. The key element in our analysis consists in showing that the 'typicality' regions are essentially disjoint. This, together with the above argument for the single message case will yield the desired result.

First, note that if $C_B$ achieves a low error probability on the asynchronous channel, then it must also achieve a low error probability on the synchronous channel—if we reveal $\nu$ to the decoder, the channel becomes synchronous and the error probability can't increase. In turn, if $C_B$ achieves a low error probability on the synchronous channel $Q$, then the strongly typical regions associated to the codewords must have 'small overlap.' Formally, it can be shown that there exists a subset $C'_B \subset C_B$ with the following properties:

a. $\log |C'_B| = \log |C_B| - o(d_B)$;
b. Every $c(m) \in C'_B$ has type $P_B$;
c. For each $c(m) \in C'_B$ and each $\hat{Q} \approx Q$, there exists $\hat{S}(m, \hat{Q})$, a subset of the output sequences that induce the joint type $P_B\hat{Q}$ with $c(m)$. For fixed $\hat{Q}$, the sets $\hat{S}(m, \hat{Q})$ are disjoint across messages, and their union has at least $|C'_B|^{2d_B H(Y|X) - o(d_B)}$ sequences.

Properties a and b are technical constraints that simplify the proofs. Property c formally captures the notion that because $C_B$ achieves low error probability on the synchronous channel, the strong typicality regions cannot overlap much. Note that for a fixed $c(m)$, the set of output strings that induce the joint type $P_B\hat{Q}$ with $c(m)$ has size $\approx d_B H(Y|X) - o(d_B)$. Therefore, c says that the size of the union is essentially maximal.

We first show how (15) follows from the existence of a subset of codewords $C'_B \subset C_B$ possessing properties a-c. To do this, we essentially mimic our argument for the case where the decoder knows $c(m)$.

For the single message case, let us parse the output sequence into blocks of size $d_B$.

As we saw above, the probability that one pure noise block induces the joint type $P_B\hat{Q}$ with $c(m)$ is $\approx 2^{-d_B D(X|Y)}$. Because now the decoder does not know which codeword was sent, the MAP decoder fails with probability at least $\frac{1}{2^{-d_B D(X|Y)}}$. Here, note that the decoder does not know which codeword was sent, the MAP decoder fails with probability at least $\frac{1}{2^{-d_B D(X|Y)}}$. But now, conditioning on the output, the MAP decoder fails with probability at least $\frac{1}{2^{-d_B D(X|Y)}}$. This shows that $d_B = \Omega(B)$. Assuming that $d_B = \Omega(B)$. Therefore, the MAP decoder fails with probability at least $\frac{1}{2^{-d_B D(X|Y)}}$. And this implies (15), completing the proof of the converse for Theorem 1.

We now prove the existence of a subset $C'_B \subset C_B$ possessing properties a-c.

Because $P_B$ is defined as the type with the most codewords from $C_B$, and the number of types is polynomial in $d_B$, we deduce that the code $C'_B$ is defined as the set of codewords in $C_B$ with type $P_B$ satisfies properties a and b. The code $C'_B$ is defined as the largest subset of $C_B$ such that there exists a corresponding deterministic decoder with the property that the maximum error probability when $C'_B$ is used over the synchronous channel $Q$ is at most $2^{-d_B/\log \log d_B}$. So defined, $C'_B$ satisfies properties b and c. Property b is clear. To see that property c holds, let $D(c(m))$ denote the decoding region associated with codeword $c(m) \in C'_B$—the decoding regions are well-defined and disjoint because the decoder is deterministic. Let $P_m$ denote the probability distribution of the output of the synchronous channel $Q$ when $c(m)$ is the channel input. Also, for a given $Q \approx Q$, let $T_Q(c(m))$ denote the set of channel outputs inducing the joint type $P_B\hat{Q}$ with $c(m)$. Then,

$$P_m(T_Q(c(m)) \cap D(c(m))) \geq P_m(T_Q(c(m)) - 2^{-d_B/\log \log d_B} \geq (1 - o(1))P_m(T_Q(c(m))),$$

where the first inequality follows from the definition of $D(c(m))$ and the second inequality follows because $Q \approx Q$ and $d_B = \Omega(B)$ implies that for all sufficiently large $B$, $P_m(T_Q(c(m))) \geq 2^{-d_B/\log \log d_B}$. Thus, conditioned on sending $c(m)$ and conditioned on the output landing within $T_Q(c(m))$, the probability of $D(c(m))$ is still at least $1 - o(1)$. Finally, this conditional distribution is uniform over $T_Q(c(m)))$, so property c is satisfied by choosing $S(m, \hat{Q}) = D(c(m)) \cap T_Q(c(m))$.

It remains to verify property a, i.e., that there exists a large subset of $C'_B$ such that there exists a corresponding (deterministic) decoder with the property that the maximum error probability is at most $2^{-d_B/\log \log d_B}$. This follows immediately from Corollary 1.9, p. 107, since $C_B$, and therefore $C'_B$, has a small maximum error probability.

There is a minor technicality in that Corollary 1.9 from only explicitly shows that a large subset exists if we move from one constant probability of error $\epsilon'$ to a smaller constant $\epsilon''$, i.e., the corollary does not allow $\epsilon''$ to depend on the blocklength $d_B$, while we require $\epsilon'' = 2^{-d_B/\log \log d_B}$. However, it can easily be verified that the arguments used to prove Corollary 1.9, p. 107 remain valid even if we choose $\epsilon'' = 2^{-d_B/\log \log d_B}$. This concludes the proof.

Converses of Theorems 14. The converse for Theorem 14 is very similar to the converse for Theorem 1, with a few minor differences due to the fact that $d_B$ is no longer subexponential in $B$. We now point to these differences.

First, to prove (14) we also use (16), but now in combination with the fact that the number of codeword types grows not

Because $Q$ is a discrete memoryless channel.
only polynomially with \( d_B \), but actually grows polynomially with \( B \), because of the positive rate constraint. This is because even if the delay \( d_B \) is exponential in \( B \), there can be at most \( O(B) \) codeword positions with non-zero cost symbols.

Second, concerning the MAP decoder after parsing the entire output sequence in blocks of size \( d_B \). When \( d_B \) grows exponentially with \( B \), the dot-equalities used to analyze the MAP decoding error are no longer valid because dot equalities are defined with respect to \( B \) rather than \( d_B \).

However, we can easily get around this problem by taking a modification, the dot equalities are defined with respect to \( B \) rather than \( d_B \). We proceed as in the subexponential delay case. This upper bound is obtained by choosing the input distribution to maximize the second term in the minimum of \( f(x) \).

To prove that this upper bound can be achieved, choose \( X \) to have a distribution with probability \( p \) to be * and probability \( 1 - p \) to be \( a^* \), with \( p \to 1 \). The first term in the min approaches:

\[
\max_x \frac{f(x)}{k(x)}
\]

by Theorem 3 of [6]. 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.

REFERENCES

[1] T.M. Cover and A. Thomas, Elements of information theory, Wiley, New York, 2006.
[2] I. Csiszár and J. Körner, Information theory: Coding theorems for discrete memoryless systems, Academic Press, New York, 1981.
[3] R. G. Gallager, Energy limited channels: Coding, multiaccess, and spread spectrum, report, M.I.T., Laboratory for Information and Decision Systems, 1987.
[4] A. Tchamkerten, V. Chandar, and G.W. Wornell, Basic limits on asynchronous communication over memoryless channels, to be submitted for publication.
[5] ———, Communication under strong asynchronism, IEEE Trans. Inform. Th. 55 (2009), no. 10, 4508–4528.
[6] S. Verdú, On capacity per unit cost, IEEE Trans. Inform. Th. 36 (1990), no. 5, 1019–1030.