Asynchronous communication: exact synchronization, universality and dispersion

Yury Polyanskiy

Abstract

Recently Tchamkerten, Chandar and Wornell proposed a novel variation of the problem of joint synchronization and error-correction. This paper considers a strengthened formulation that requires the decoder to estimate both the message and the location of the codeword exactly. Such a scheme allows for transmitting data bits in the synchronization phase of the communication, thereby improving bandwidth and energy efficiencies. It is shown that the capacity region remains unchanged under the exact synchronization requirement. Furthermore, asynchronous capacity can be achieved by universal (channel independent) codes. Comparisons with earlier results on another (delay compensated) definition of rate are made. The finite blocklength regime is investigated and it is demonstrated that even for moderate blocklengths, it is possible to construct capacity-achieving codes that tolerate exponential level of asynchronism and experience only a rather small loss in rate compared to the perfectly synchronized setting; in particular, the channel dispersion does not suffer any degradation due to asynchronism. For the binary symmetric channel a translation (coset) of a good linear code is shown to achieve the capacity-synchronization tradeoff.

Index Terms

Shannon theory, channel capacity, channel coding, universal codes, asynchronous communication, synchronization, strong converse, non-asymptotic analysis, finite blocklength, discrete memoryless channels

The author is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139 USA. e-mail: yp@mit.edu. This work supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. This work was presented in part at the 2012 Conference on Information Sciences and Systems (CISS), Princeton University, Princeton, NJ.

I. Introduction

The traditional approach to the problem of reliable communication in the presence of noise typically assumes that the decoder has access to a corrupted version of the original waveform with the beginning and the end of the waveform being perfectly known. In such a setting for sufficiently long blocklengths modern sparse graph codes achieve almost the best possible error correction. It is natural, therefore, to revisit other sources of suboptimality in a communication system. One such overhead is introduced by the traditional frame synchronization methods [1]–[3] employing periodical pilot signals that consume both energy and bandwidth. The focus of this paper is the problem of performing the error-correction and synchronization jointly.

Classically, several approaches going beyond pilot-based methods were investigated. One of the earliest ideas is comma-free encoding [4] that allows one to recover synchronization in a data stream composed of back-to-back transmitted codewords in the absence of channel noise. In [5] it was shown that a coset of a good linear code will have the comma-free property. Extension to situation with channel noise was considered in [6], where again it was shown that cosets of certain linear codes suffice. Another line of work focused on coding [7] and fundamental limits [8] for communication in the presence of random insertions/deletions. In an even earlier work Gallager [9] shows that for this setting a good (convolutional) code scrambled by a pseudo-random sequence yields an excellent synchronization performance. In the context of multiple-access channels, treatments of both the frame-asynchronism [10]–[12] and symbol-asynchronism [13] focused on the case when the relative time offsets between the users are perfectly known at the decoder (or remain constant across multiple transmissions, which makes them reliably learnable at the decoder).

This paper considers the problem of initial acquisition of synchronization. Namely, the actual noisy transmission is assumed to be preceded by a (random length) sequence of background noise. The goal of the decoder is to detect the precise moment when the message starts as well as to decode the content of the initial frame. The motivation is to optimize energy and bandwidth efficiency of modern systems by inserting data bits into the synchronization phase of communication. In practice, these bits might be used for implementing multiple-access to a shared wideband medium, e.g. the receiver may be synchronizing with a frequency-hopping pattern corresponding to a specific user.

Such a single shot model of asynchronism has recently been proposed by Tchamkerten, Chandar and Wornell [14], who were motivated in part by the sensor networks in which nodes exchange data very infrequently (thus, making constant channel-tracking impractical). Subsequent work [15], [16] demonstrated significant advantages in going beyond the conventional pilot-based approach. The focus of [14]–[16] was on recovering the message only, whereas this paper considers both the message and the timing. In the context of initial synchronization such extension appears to be vital.

Mathematically, the formulation of [14] is a generalization of the change point detection problem [17], close in spirit to the so called "detection and isolation" problem introduced in [18], except that in the latter the set of distributions that the original one can switch to is pre-specified whereas [14] allows for an optimal codebook design.

Here we show that the requirement of timing recovery does not change the capacity region compared to the one reported in [16] (for the special case when cost of each symbol is 1). For binary symmetric channel this is achieved by a random coset of a good linear code, which is in agreement with the classical findings discussed above. In addition show that sequences of universal codes exist that achieve all points inside the capacity region simultaneously for all channels. Finally, we investigate the results in the regime of finite blocklength. In particular, we demonstrate that even for short blocklengths it is possible to combat a gigantic (exponential) asynchronism while achieving essentially the same performance as for the synchronous setting: namely, the channel dispersion [19] is unchanged. This illustrates that communication systems, investing significant bandwidth in pilots, may be operating far from optimality.

The organization of the paper is as follows. Section II defines the problem formally. Section III contains the asymptotic results on the capacity and universality, and comparisons with the results in [14]–[16]. Section IV presents the finite blocklength results and draws conclusions on channel dispersion. With the exception of the non-asymptotic achievability bound in Theorem 4 the discussion focuses on discrete memoryless channels (DMCs).

II. Problem formulation and notation

Consider a DMC with stochastic matrix $W : \mathcal{X} \to \mathcal{Y}$ and a distinguished symbol $\star \in \mathcal{X}$. We define its blocklength n extension as

$$W^n(y^n|x^n) = \prod_{j=1}^{n} W(y_j|x_j)\,. \tag{1}$$

Given a number $A_n \geq n$ we define an asynchronous random transformation, denoted (W^n, A_n), as follows:

- input space is $\mathcal{X}^n \stackrel{\triangle}{=} \{(x_1, \ldots, x_n) : x_i \in \mathcal{X}, i = 1, \ldots, n\}$
- output space is $\mathcal{Y}^{A_n}$
- the transformation acts as follows:

$$P_{Y^{A_n}|X^n}(\cdot|\cdot) = \sum_t P_\nu(t) P_{Y^{A_n}|X^n,\nu}(\cdot|\cdot, t)\,,$$

  where ν is a random variable uniformly distributed on $\{1, \ldots, A_n - n + 1\}$ and

$$P_{Y^{A_n}|X^n,\nu}(y^{A_n}|x^n, t) = \prod_{j<t, j\geq t+n} W(y_j|\star) \prod_{t\leq j<t+n} W(y_j|x_j)\,. \tag{2}$$

Definition 1: An M-code for the random transformation (W^n, A_n) is a triplet

- An encoder function $f : \{1, \ldots, M\} \to \mathcal{X}^n$
- A stopping time $n \leq \tau \leq A_n - n + 1$ of the filtration generated by $\{Y_j, j = 1, \ldots, A_n\}$. For convenience, we set

$$\hat{\nu} \stackrel{\triangle}{=} \tau - n + 1\,,$$

  which marks the decoder's estimate of ν.
- A decoder function $g : \mathcal{Y}^{\hat{\nu}+n-1} \to \{1, \ldots, M\}$.

Given an M-code we construct a probability space $(\mathcal{W}, X^n, Y^{A_n}, \hat{\mathcal{W}})$ and distribution $\mathbb{P}$ on it by taking $\mathcal{W}$ – uniform on $\{1, \ldots, M\}$, $\hat{\mathcal{W}} = g(Y^{\hat{\nu}+n-1})$ and then chaining all random transformations according to the directed graphical model:

$$\begin{array}{ccccccc} & & & & \nu & & \\ & & & & \searrow & & \\ \mathcal{W} & \xrightarrow{f} & X^n & \longrightarrow & Y^{A_n} & \xrightarrow{\hat{\nu},g} & \hat{\mathcal{W}} \end{array} \tag{3}$$

The code is said to be an (M, ϵ) code if its probability of error does not exceed ϵ. In this paper we consider three definitions of probability of error (in the order of decreasing strength):

$$\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W}, \hat{\nu} = \nu] \geq 1 - \epsilon \tag{4}$$

$$\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W}, \hat{\nu} \leq \nu] \geq 1 - \epsilon \tag{5}$$

$$\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W}, \hat{\nu} \leq \nu + L_n] \geq 1 - \epsilon\,, \qquad L_n = \exp\{o(n)\} \tag{6}$$

An (M, ϵ)-code under (4) is required to decode the message $\mathcal{W}$ and synchronize, under (5) is required to decode the message $\mathcal{W}$ with no additional delay (lookahead past the end of the codeword), while (6) allows for a subexponential delay L_n. The criterion (6) was considered in [16] and only introduced here for the purpose of comparison.

One of the main results of this paper is that (rate-wise) it does not matter which one of the three criteria is used and whether ϵ is held fixed or is asymptotically vanishing. Thus, neither enabling the lookahead nor permitting the early decision change the capacity, which may appear somewhat surprising considering that the commonly used pilot-based synchronization scheme of Massey [20] does in fact require some lookahead past the pilot end (called deferred-decision in [2]).

Definition 2: A pair $(R, \mathcal{A})$ is called ϵ-achievable if there exist sequences of numbers $A_n \geq n$ and $M_n \geq 2$ satisfying

$$\liminf_{n\to\infty} \frac{1}{n} \log A_n \geq \mathcal{A}\,, \tag{7}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log M_n \geq R \tag{8}$$

and a sequence of (M_n, ϵ) codes for random transformations (W^n, A_n). The asynchronous ϵ-capacity at asynchronism $\mathcal{A}$ is defined as

$$C_\epsilon(\mathcal{A}) \stackrel{\triangle}{=} \sup\{R : (R, \mathcal{A}) \text{ is } \epsilon\text{-achievable}\}\,.$$

The asynchronous capacity at asynchronism $\mathcal{A}$ is defined as

$$C(\mathcal{A}) \stackrel{\triangle}{=} \lim_{\epsilon\to 0} C_\epsilon(\mathcal{A})\,.$$

The ϵ-synchronization threshold $\mathcal{A}_{\circ,\epsilon}$ is defined as

$$\mathcal{A}_{\circ,\epsilon} \stackrel{\triangle}{=} \sup\{\mathcal{A} : (0, \mathcal{A}) \text{ is } \epsilon\text{-achievable}\}$$

and the synchronization threshold is

$$\mathcal{A}_\circ \stackrel{\triangle}{=} \lim_{\epsilon\to 0} \mathcal{A}_{\circ,\epsilon}\,.$$

Remark: Note that $(0, \mathcal{A})$ is ϵ-achievable if and only if there exist a sequence of $(n, 2, \epsilon)$ codes for random transformations $(W^n, 2^{n\mathcal{A}+o(n)})$.

The main difference with the model studied in [14], [15] is that the definition of rate there was

$$\tilde{R} \stackrel{\triangle}{=} \frac{\log M}{\mathbb{E}\left[|\hat{\nu} - \nu + n|^+\right]} \tag{9}$$

and correspondingly the error event was defined as just $\{\hat{\mathcal{W}} \neq \mathcal{W}\}$. With such a (delay compensated) definition of rate, one defines the capacity $\tilde{C}(\mathcal{A})$ in exactly the same manner as $C(\mathcal{A})$; the key results of [14], [15] provide upper and lower bounds on $\tilde{C}(\mathcal{A})$ (but not $\tilde{C}_\epsilon(\mathcal{A})$). The definition (9) was chosen, perhaps, to model the situation when one wants to assess the minimal number of channel uses (per data bit) that the channel remains under the scrutiny of the decoder, whereas our definition

$$R \stackrel{\triangle}{=} \frac{\log M}{n} \tag{10}$$

serves the purpose of studying the minimal number of channel uses (per data bit) that the channel remains occupied by the transmitter, while the delay constraint is disentangled from the rate definition by the condition (5). With such definitions, our model can be interpreted as the problem of communicating both the data $\mathcal{W}$ and the state ν as in [21], except that the state is no longer a realization of the discrete memoryless process and it enters the channel law (W^n, A_n) in a different way.

The notation in this paper follows that of [22] and [19, Section IV.A], in particular, $D(P||Q)$ denotes the relative entropy between distributions P and Q; $W_x(\cdot) = W(\cdot|x)$; for a distribution P on $\mathcal{X}$ a distribution PW on $\mathcal{Y}$ is defined as $PW(y) = \sum_x W(y|x)P(x)$; we agree to identify distribution Q on $\mathcal{Y}$ with a stochastic kernel $Q : \mathcal{X} \to \mathcal{Y}$ which is constant on $\mathcal{X}$, so under this agreement $PW_x = W_x$; and $I(P, W)$ is a mutual information between $X \sim P$ and $Y \sim PW$ and coupled via $P_{Y|X} = W$: $I(P, W) = D(W||PW|P)$. We denote by P^n the product distribution on $\mathcal{X}^n$ and similarly for $\mathcal{Y}^n$. We adopt the definitions of [22, Chapter 2] for the concepts of an n-type, a typical set T_P, V-shells $T_V(x)$, etc. Additionally, we agree for any set S of stochastic

matrices $V : \mathcal{X} \to \mathcal{Y}$ to denote

$$T_S(x) = \bigcup_{V \in S} T_V(x)\,. \tag{11}$$

The spaces of probability measures on $\mathcal{X}$ and $\mathcal{Y}$, and stochastic matrices $V : \mathcal{X} \to \mathcal{Y}$ are given the topology inherited from the canonical identification with the convex compact subsets of the respective finite dimensional Euclidean spaces.

III. Asymptotic results

We summarize the previously known results:

Theorem 1 ([14], [16], [23]): For any DMC W and $0 < \epsilon < 1$ we have

$$\mathcal{A}_{\epsilon,\circ} = \mathcal{A}_\circ = \max_{x \in \mathcal{X}} D(W_x||W_\star)\,. \tag{12}$$

The asynchronous capacity and ϵ-capacity of the DMC W under the probability of error criteria (5) or (6) is:

$$C(\mathcal{A}) = C_\epsilon(\mathcal{A}) = \max_{P: D(PW||W_\star) \geq \mathcal{A}} I(P, W)\,, \tag{13}$$

where the maximum is defined to be zero whenever $\mathcal{A} > \mathcal{A}_\circ$.

Remark: Results regarding (5) are not mentioned in [16] explicitly, but maybe extracted from the proofs. Similarly, the strong converse part, i.e. $C_\epsilon(\mathcal{A}) = C(\mathcal{A})$, is implicitly contained in [16]. For completeness, in Appendix A we put an alternative proof of the strong converse. The proof strategy is similar to [16], but we build upon the hypothesis-testing framework of [19], which allows us to do away with a complicated refinement of the blowing-up lemma required by [16][1].

Remark: As shown in [16, Theorem 5] the weak converse in Theorem 1 is unchanged if ν is not precisely uniform on $\exp\{n\mathcal{A}\}$ atoms but rather is "essentially" such: namely, the length ℓ_n of the optimal binary lossless compressor of ν satisfies:

$$\frac{1}{n}\ell_n \to \mathcal{A}\,,$$

where the convergence is in probability, which by a standard argument is equivalent to

$$\frac{1}{n}\log\frac{1}{P_\nu(\nu)} \to \mathcal{A}\,, \tag{14}$$

also in probability. As the argument in Appendix A demonstrates under such an assumption the strong converse continues to hold also.

[1]*Added in proof:* A recent revision [24] appears to work around the blowing-up lemma as well.

Our main asymptotic results are the following:

Theorem 2: For any DMC W and any $0 < \epsilon < 1$ the asynchronous capacity $C_\epsilon(\mathcal{A})$ under the stronger criterion (4) is given by (13). In other words, the requirement of precise timing recovery as per (4) does not incur any loss in rate.

It turns out that all rates up to capacity $C(\mathcal{A})$ can also be approached by a universal sequence of codes that does not require an apriori knowledge of W:

Theorem 3: Fix $\mathcal{A} \geq 0$, rate R and a distribution P on $\mathcal{X}$. Then there exists a sequence of codebooks and universal decoders which simultaneously achieves a vanishing probability of error (5) over all asynchronous DMCs $(W^n, \exp\{\mathcal{A}n\})$ satisfying

$$\mathcal{A} \;<\; D(PW||W_\star)\,, \tag{15}$$

$$R \;<\; I(P,W)\,. \tag{16}$$

The equivalence of Theorem 3 with the capacity expression (13) follows from

$$\max_{P:D(PW||W_\star)\geq\mathcal{A}} I(P,W) = \sup_{P:D(PW||W_\star)>\mathcal{A}} I(P,W)\,, \tag{17}$$

where supremum in the left-hand side is taken to be zero if the constraint set is empty. To show (17), note that if the maximizer P in the right-hand side is such that $D(PW||W_\star) > \mathcal{A}$ then there is nothing to prove, so assume $D(PW||W_\star) = \mathcal{A}$. If P is a local maximum of $P \mapsto D(PW||W_\star)$ then by convexity it must be a global one and in particular

$$D(W||W_\star|P) \leq D(PW||W_\star)$$

implying that $I(P,W) = D(W||W_\star|P) - D(PW||W_\star) = 0$ and both sides of (17) are zero. Otherwise, if P is not a local maximum, there must be a sequence $P_n \to P$ such that $D(P_nW||W_\star) > \mathcal{A}$, which implies the equality in (17) by continuity of $P_n \mapsto I(P_n, W)$.

A. Discussion and comparison of results

As an example of evaluating (12)-(13), consider the binary symmetric channel $BSC(\delta)$ with $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{0,1\}$, $\star = 0$ and

$$W(y|x) = \begin{cases} 1-\delta\,, & y = x \\ \delta\,, & y \neq x\,, \end{cases}$$

For such a model, computation of (12)-(13) yield

$$\mathcal{A}_\circ = d(\delta||1-\delta)\,, \tag{18}$$

$$C(d(p*\delta||1-\delta)) = h(p*\delta)-h(\delta)\,, \qquad p\in[0,\tfrac{1}{2}]\,, \tag{19}$$

where the latter is presented in parametric form and we have defined

$$d(x||y) = x\log\frac{x}{y}+(1-x)\log\frac{1-x}{1-y}\,, \tag{20}$$

$$h(x) = x\log\frac{1}{x}+(1-x)\log\frac{1}{1-x}\,, \tag{21}$$

$$p*\delta = (1-p)\delta+p(1-\delta)\,. \tag{22}$$

For the analogous case of binary erasure channel, the $C(\mathcal{A})$ equals the usual synchronous capacity C for all $\mathcal{A}>0$. Indeed, if $\mathcal{A}_\circ=\infty$ then according to (13)

$$C(\mathcal{A}) = \max_P I(P,W) \stackrel{\triangle}{=} C\,, \qquad \forall \mathcal{A}\geq 0\,, \tag{23}$$

i.e. capacity can be achieved for all exponents $\mathcal{A}\geq 0$.

Next, we compare results in Theorems 2 and 3 with the previously known Theorem 1:

- Theorem 2 proves achievability part under a more stringent condition (4). Unlike [16] (and [15]) our proof relies on analyzing the behavior of information density, a certain (super-)martingale property of which allows us to guarantee the perfect synchronization required in (4). An additional benefit is that the resulting bounds are competitive in the finite blocklength regime, as shown later in Section IV.
- The fact that requirement of perfect synchronization does not incur a penalty in rate may appear somewhat surprising. For example, as shown in [23] (which considers synchronization problem only, i.e. there is only $M=1$ message), to find the exact location of the start of the transmission one may use a shift-register generated pseudo-random sequence. However, it turns out that no preamble-based method may achieve asynchronous capacity (see the remark after (24) and [15]). Nevertheless, Theorem 2 constructs a codebook that is usable for blind synchronization without the need for preambles with favorable autocorrelation.
- Theorem 3 relies on a generalization of the packing lemma, which is used to construct a codebook having vanishing probability of error simultaneously for a class of DMCs.
- The alternative proof of the strong converse of Theorem 1 that we presented in Appendix A shows that the ϵ-capacity is unchanged even if the distribution of the start of transmission

ν is non-uniform as in (14). Our proof technique is of independent interest, as it further develops the meta-converse framework [19, Section III.E] and [25, Section 2.7], which is known to result in tight non-asymptotic bounds (see [19, Section III.J4]). It is possible that our methods would also prove useful for improving the bounds on the capacity $\tilde{C}(\mathcal{A})$ in the model (9).

It is instructive to compare results of Theorems 1 and 2 to those in [14], [15] for a delay-compensated definition of rate (9):

- In both cases the synchronization threshold is given by (12); see [14]. This is not surprising since (as remarked above) $\mathcal{A}_\circ$ is determined by the ability to communicate with $M = 2$ codewords, for which the precise definition of rate is immaterial.
- In both cases, there is a "discontinuity at $R = C$" in the sense that $C(\mathcal{A}) = C$ for all $\mathcal{A} \le \mathcal{A}_1$ with $\mathcal{A}_1 > 0$ if and only if

$$D(P_Y^*||W_\star) > 0\,,$$

  where P_Y^* denotes the unique capacity achieving output distribution. However, the precise value of this critical exponent $\mathcal{A}_1$ is unknown for the model (9) even for the BSC, whereas in the model (10) we always have

$$\mathcal{A}_1 = D(P_Y^*||W_\star)\,. \tag{24}$$

- In both cases, for a certain natural class of synchronization schemes based on preambles, see [15, Definition 3], we have $\mathcal{A}_1 = 0$, which prevents achieving capacity with positive asynchronism exponent. For the model (9) this is shown in [15, Corollary 3], while for the model (10) this is simply trivial: to combat a positive asynchronism exponent one would require preamble of the size δn, but this penalizes the rate to be at most $C - \delta$.
- According to [15] there exist channels (and BSC is one of them – see below) for which the capacity $\tilde{C}(\mathcal{A}) = 0$ for some range of $\mathcal{A} < \mathcal{A}_\circ$. In such regime there exist codes reliably sending $M = \exp\{nR\}$ codewords, but the rate $\tilde{R}$, as defined in (9), remains zero. This strange behavior, called "discontinuity at $R = 0$" in [15, Corollary 2] does not occur in the definition of rate (10): the capacity is positive for all $\mathcal{A} < \mathcal{A}_\circ$.
- Somewhat counter-intuitively although in our model we impose a seemingly strong condition $\{\hat{\nu} \le \nu\}$ absent in [14], [15], it turns out that the capacity vs. asynchronous exponent region

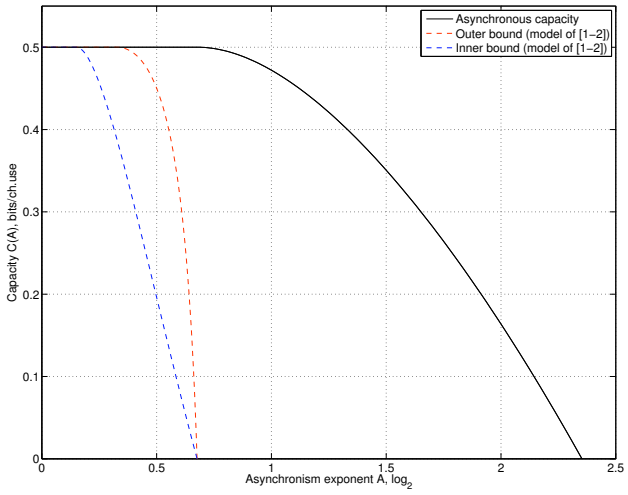

Fig. 1. $BSC(0.11)$: The asynchronous capacity (19) compared with inner-outer bounds of [15] for the different model (9).

is larger. This is explained by noticing that if $\hat{\nu} > \nu$ then one typically has $\hat{\nu} = \nu + \exp\{n\epsilon\}$. Thus in the model (9), to avoid significant penalty in rate the occurrence of $\hat{\nu} > \nu$ should happen with exponentially small probability.

Additionally, [26] considers the definition of rate as in (10) but models asynchronism differently and restricts the decoders to operate solely on the basis of each observed n-block. Their region of rate vs. false alarm error-exponent coincides with the region (13) of rate vs. asynchronism exponent; see [26, Theorem 1]. This is explained by noticing that the false alarm is what governs the level of asynchronism that the code is capable of tolerating.

To illustrate these points, in Fig. 1 we compare the region (13) with inner (achievability) and outer (converse) bounds found in [15, Theorem 2 and Theorem 3], respectively, which for the case of the $BSC(\delta)$ can be shown to be (in parametric form)

$$\mathcal{A} = d(q||\delta), \quad \tilde{C}_{in}(\mathcal{A}) = h(p * \delta) - h(\delta) \tag{25}$$

$$\mathcal{A} = \frac{d(p * \delta||1-\delta)d(\frac{1}{2}||\delta)}{d(p * \delta||1-\delta) + pd(\frac{1}{2}||\delta)}, \quad \tilde{C}_{out}(\mathcal{A}) = h(p * \delta) - h(\delta) \tag{26}$$

where parameter runs over $p \in [0, \frac{1}{2}]$ and in (25) q solves

$$d(q||p * \delta) = d(q||\delta).$$

Note that according to the $\tilde{C}_{out}$ bound the capacity in the model (9) is zero between $d(\frac{1}{2}||\delta)$ and $\mathcal{A}_\circ = d(1-\delta||\delta)$. This demonstrates the above mentioned discontinuity at $R = 0$ for the

BSC and therefore closes the open question mentioned after [15, Corollary 2].

B. Proof of Theorem 2

The main problem in achieving a good error-correction performance in the presence of asynchronism is the ability to resolve partially aligned codewords. For example, suppose that a codeword $x \in \mathcal{X}^n$ is being transmitted. Then, if there is a k-symbol misalignment, $0 \le k < n$, the decoder observes outputs effectively generated by a shifted codeword $x^{\star k}$:

$$x^{\star k} \stackrel{\triangle}{=} (\underbrace{\star, \ldots, \star}_{k}, x_1, \ldots, x_{n-k}) \in \mathcal{X}^n \,. \tag{27}$$

As will be shown shortly, in asynchronous communication achieving a small probability of error in the sense of (5) requires constructing a codebook in which any pair of distinct codewords $c, \bar{c} \in \mathcal{X}^n$ are far apart, and so are $c^{\star k}$ and $\bar{c}$. As illustrated by [15, Theorem 2] and [16, Theorem 1], the existence of such codebooks follows immediately from a random coding argument, since $c^{\star k}$ and $\bar{c}$ are independent. Below we materialize this intuition into a finite-blocklength achievability result (Theorem 4) and a universal packing lemma (Lemma 7).

Achieving a small probability of error in the stronger sense of (4), however, requires constructing a codebook in which $c^{\star k}$ is far away from c itself. Because of strong dependence between $c^{\star k}$ and c this presents a new type of difficulty. Interestingly, however, in memoryless channels this dependence may still be controlled. For example, for the BSC the quality of distinguishing c from $\bar{c}$ depends, essentially, on the Hamming distance $|c - \bar{c}|$ only. If $c, \bar{c} \in \{0,1\}^n$ are selected uniformly, then evidently $|c - \bar{c}|$ and $|c - c^{\star k}|$, $k \ge 1$ have identical (binomial) distribution. Thus, on average c is distinguishable from both $\bar{c}$ and $c^{\star k}$ with small probability of error, resulting in correct synchronization and message decoding as required by (4). For a general DMC, this observation is the content of Lemma 6 below.

We proceed to formal analysis. First, we show how to achieve small probability of error in (5). The following bound applies to general (non-memoryless, non-discrete) channels $P_{Y^n|X^n}$ and will also be used for finite blocklength evaluation in Section IV:

Theorem 4: Consider an arbitrary random transformation $P_{Y^n|X^n} : \mathcal{X}^n \to \mathcal{Y}^n$. Then for any $\gamma \ge 0$ and any input distribution P_{X^n} on $\mathcal{X}^n$ there exists an M-code for the random

transformation $(P_{Y^n|X^n}, A)$ satisfying

$$\begin{aligned}
&\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W}, \nu - n < \hat{\nu} \leq \nu] \\
&\quad \geq \quad P[i(X^n;Y^n) > \gamma] - nM\exp\{-\gamma\} - \mathbb{E}\left[\exp\{-|r(Y^n) - \log A|^+\}\right], \qquad (28)
\end{aligned}$$

where P denotes probability with respect to the distribution $P_{X^nY^n}(x,y) = P_{Y^n|X^n}(y|x)P_{X^n}(x)$, $\mathbb{E}$ is the expectation with respect to P and we also defined

$$r(y^n) \stackrel{\triangle}{=} \log\frac{P_{Y^n}(y^n)}{W_\star(y^n)} \qquad (29)$$

$$i(x^n;y^n) \stackrel{\triangle}{=} \log\frac{P_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)}. \qquad (30)$$

Remark: As the proof demonstrates, the bound holds for an arbitrary (not necessarily uniform) distribution of ν on $\{1,\ldots,A\}$.

Proof: Intuitively, good asynchronous decoder will analyze sequentially each n-block of outputs y^n and try to determine whether it was generated by one of the codewords or the background noise. The decoding stops once y^n is "close" to one of the codewords c_i while also being "far" from the typical $W_\star$-noise. Naturally, then for a given codebook $\{c_1,\ldots,c_M\}, c_i \in \mathcal{X}^n$ we define the asynchronous decoder as follows:

$$\hat{\nu} \stackrel{\triangle}{=} \inf\{t \geq 1 : \exists j : r(Y_t^{t+n-1}) \geq \gamma_1, i(c_j; Y_t^{t+n-1}) > \gamma\}, \qquad (31)$$

$$\hat{\mathcal{W}} \stackrel{\triangle}{=} \min\{j : i(c_j; Y_{\hat{\nu}}^{\hat{\nu}+n-1}) > \gamma\}, \qquad (32)$$

where γ_1 is a constant to be chosen later. We now replace each c_j with a random codeword C_j. The elements of the codebook $\{C_j, j = 1,\ldots M\}$ are generated independently of each other with distribution P_{X^n}. We proceed to upper bound the probability of the error event

$$E \stackrel{\triangle}{=} \{\hat{\mathcal{W}} \neq 1\} \cup \{\hat{\nu} > \nu\} \cup \{\hat{\nu} \leq \nu - n\}. \qquad (33)$$

For this computation we assume without loss of generality that $\mathcal{W} = 1$. Then, we have

$$\begin{aligned}
\mathbb{P}[E] &\leq \mathbb{P}[E, i(C_1; Y_\nu^{\nu+n-1}) > \gamma] + \mathbb{P}[i(C_1; Y_\nu^{\nu+n-1}) \leq \gamma] && (34)\\
&= \mathbb{P}[E, i(C_1; Y_\nu^{\nu+n-1}) > \gamma] + P[i(X^n;Y^n) \leq \gamma] && (35)\\
&\leq \mathbb{P}[\hat{\nu} > \nu, i(C_1; Y_\nu^{\nu+n-1}) > \gamma] + \mathbb{P}[\hat{\nu} \leq \nu - n] && (36)\\
&\quad + \sum_{k=0}^{n-1} \mathbb{P}[\hat{\mathcal{W}} \neq 1, \hat{\nu} = \nu - k, i(C_1; Y_\nu^{\nu-n+1}) > \gamma] + P[i(X^n;Y^n) \leq \gamma] && (37)
\end{aligned}$$

where in (35) we rewrote the second term via the probability P from (28) and (37) is as in (74).

First term is bounded as follows:

$$\begin{aligned}\mathbb{P}[\hat{\nu} > \nu, i(C_1; Y_\nu^{\nu+n-1}) > \gamma] &\leq \mathbb{P}[r(Y_\nu^{\nu+n-1}) < \gamma_1] &(38)\\ &= P_{Y^n}[r(Y^n) < \gamma_1]\,, &(39)\end{aligned}$$

which follows by noticing that $Y_\nu^{\nu+n-1}$ is distributed precisely as P_{Y^n}. The second term is also handled easily:

$$\begin{aligned}\mathbb{P}[\hat{\nu} \leq \nu - n] &\leq \mathbb{P}[\exists t < \nu - n : r(Y_t^{t+n-1}) \geq \gamma_1] &(40)\\ &\leq A_n W_\star^n[r(Y^n) \geq \gamma_1]\,, &(41)\end{aligned}$$

since for $t < \nu - n$ we have $Y_t^{t+n-1} \sim W_\star^n$.

For the third term in (37) we have

$$\begin{aligned}\sum_{k=0}^{n-1} \mathbb{P}[\hat{\mathcal{W}} \neq 1, \hat{\nu} = \nu - k, i(C_1; Y_\nu^{\nu-n+1}) > \gamma] &\leq \sum_{k=0}^{n-1}\sum_{j=2}^{M} \mathbb{P}[i(C_j, Y_{\nu-k}^{\nu-k+n-1} \geq \gamma] &(42)\\ &\leq nM \exp\{-\gamma\}\,, &(43)\end{aligned}$$

where (42) follows simply by the fact that under $\{i(C_1; Y_\nu^{\nu-n+1}) > \gamma\}$ only makes an error if for another codeword and some time shift $\nu - n < t \leq \nu$ we have $i(C_j; Y_t^{t+n-1}) > \gamma$, and (43) follows because C_j and Y_t^{t+n-1} are independent and thus the standard property of information density applies:

$$P\left[\log \frac{P_{AB}(A, \bar{B})}{P_A(A)P_B(\bar{B})} \geq \gamma\right] \leq \exp\{-\gamma\}\,,$$

whenever A and $\bar{B}$ are independent with $A \sim P_A$ (or, by symmetry, if $\bar{B} \sim P_B$).

Finally, summing all terms together we get

$$\mathbb{P}[E] \leq P_{Y^n}[r(Y^n) < \gamma_1] + A_n W_\star^n[r(Y^n) \geq \gamma_1] + P[i(X^n; Y^n) \leq \gamma] + nM \exp\{-\gamma\}\,. \quad (44)$$

Choosing now $\gamma_1 = \log A$ and applying the identity [19, (69)] to the first two terms in (44) we get (28). Optimality of the choice $\gamma_1 = \log A$ follows the same argument as in the discussion after [19, Lemma 19]. ■

Next, we prove the achievability bound for a stronger criterion (4). This result is restricted to DMC.

Theorem 5: Let $W : \mathcal{X} \to \mathcal{Y}$ be a DMC. Let $\tau > 0$ be arbitrary. Then, there exists $E > 0$ such that for any input distribution P on $\mathcal{X}$, any sufficiently large n, any $\gamma \geq n\tau$ and any M and A there exists an M-code for the asynchronous random transformation (W^n, A) satisfying

$$\begin{aligned}
&\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W}, \hat{\nu} \neq \nu] \\
&\geq \quad P[i(X^n; Y^n) > \gamma] - nM \exp\{-\gamma\} \\
&\qquad -\mathbb{E}\left[\exp\{-|r(Y^n) - \log A|^+\}\right] - (n-1)\exp\{-nE\}\,, \qquad (45)
\end{aligned}$$

where P denotes probability with respect to the distribution $P_{X^nY^n}(x, y) = W^n(y^n|x^n)P^n(x^n)$, $\mathbb{E}$ is the expectation with respect to P and $r(\cdot)$, $i(\cdot;\cdot)$ are defined by (29), (30)

Proof: By following the random-coding construction in Theorem 4 with $P_{X^n} = P^n$ we show that the probability of event $\{\hat{\mathcal{W}} = \mathcal{W}, \nu - n < \hat{\nu} \leq \nu\}$ averaged over all codebooks is lower-bounded by the left-hand side of (28). This takes care of the first three terms in (45). Thus, we only need to prove that the decoder (31)-(32) also achieves

$$\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W}, \nu - n < \hat{\nu} < \nu] \leq (n-1)\exp\{-nE\} \qquad (46)$$

for all sufficiently large n and some $E > 0$ (when averaged over the random codebook $\{C_1, \ldots, C_M\}$). As in (37) by symmetry we may condition on $\mathcal{W} = 1$. Note that for any $1 \leq k < n$ we have

$$\{\hat{\mathcal{W}} = 1, \hat{\nu} = \nu - k\} \subseteq \{i(C_1; Y_{\nu-k}^{\nu-k+n-1}) \geq \gamma\}\,.$$

Since $\gamma \geq n\tau$, by Lemma 6 (to follow next) we have

$$\mathbb{P}[i(C_1; Y_{\nu-k}^{\nu-k+n-1}) \geq \gamma] \leq \exp\{-nE\}\,. \qquad (47)$$

Then (47) and the union bound imply (46) and (45). ■

Lemma 6: Fix DMC $W : \mathcal{X} \to \mathcal{Y}$, input distribution P on $\mathcal{X}$ and let

$$i(a^n; b^n) = \sum_{j=1}^{n} \log \frac{W(b_j|a_j)}{PW(b_j)}$$

or $i(a^n; b^n) = -\infty$ if $W(b_j|a_j) = 0$ for at least one j. Let $X_j = \star$ for $-n+1 \leq j \leq 0$ and X_j be i.i.d. with law P for $1 \leq j \leq n$. Let Y_{-n+1}^n be the result of passing X_{-n+1}^n over the DMC W. Then for any $\tau > 0$ there exists $E > 0$ such that for all sufficiently large n and all $1 \leq k \leq n$ we have

$$\mathbb{P}[i(X_1^n; Y_{1-k}^{n-k}) > n\tau] \leq \exp\{-nE\}\,. \qquad (48)$$

Remark: Thus using a misaligned output Y_{1-k}^{n-k} instead of Y_1^n results in a log-likelihood ratio that is almost always non-positive.

Proof: Let $W^r : \mathcal{Y} \to \mathcal{X}$ be the reverse DMC defined as

$$W^r(x|y) = \frac{W(y|x)P(x)}{PW(y)}\,.$$

Note that for $1 \le j \le n$ we have

$$T_j = \log \frac{W(Y_{j-k}|X_j)}{PW(Y_{j-k})} = \log \frac{W^r(X_j|Y_{j-k})}{P(X_j)}\,.$$

Therefore, by independence of X_j's we have

$$\mathbb{E}\left[T_j | X_{-n+1}^{j-1}, Y_{-n+1}^{j-1}\right] = -D(P||W^r_{Y_{j-k}}) \le 0\,,$$

where as usual W_b^r denotes a distribution $W^r(\cdot|b)$ on $\mathcal{X}$. Thus, the cumulative sums $\sum_{j=1}^m T_j$ form a super-martingale. Since

$$i(X_1^n; Y_{1-k}^{n-k}) = \sum_{j=1}^n T_j$$

estimate (48) follows from Azuma's inequality once we ensure that the T_j are bounded. To that end notice that

$$T_j \le \log \frac{1}{p_{min}}\,,$$

where p_{min} is a minimal non-zero value of $P(\cdot)$. If $\min_{x,y} W(y|x) > 0$ then the lower bound follows similarly. If however, stochastic matrix W has zeros the T_j and $D(P||W_y^r)$ may be infinite. This can be fixed as follows. By finiteness of $\mathcal{Y}$ we can always find a very large $\theta > 0$ such that for all $y \in \mathcal{Y}$

$$\begin{aligned} -D(P||W_y^r) &\le \sum_{x\in\mathcal{X}} P(x) \max\left(-\theta, \log \frac{W^r(x|y)}{P(x)}\right) && (49)\\ &\le 0\,. && (50)\end{aligned}$$

Then, since

$$i(X_1^n; Y_{1-k}^{n-k}) \le \sum_{j=1}^n \max(-\theta, T_j)$$

the (48) follows by applying Azuma's inequality to a bounded difference super-martingale $\sum_{j=1}^n \max(-\theta, T_j)$. ∎

Finally, we may put all the pieces together:

Proof of Theorem 2: : By (17), to prove achievability of (13) it is enough to show that for every triple $P, R, \mathcal{A}$ such that (15)-(16) hold, the pair $(R, \mathcal{A})$ is achievable with a vanishing probability of error in the sense of (4). To that end, we apply Theorem 5 with $M = \exp\{nR\}$, $A = \exp\{n\mathcal{A}\}$, $\gamma = nR + n\delta$ where $\delta > 0$ is such that

$$R + \delta \;<\; I(P, W) \tag{51}$$

$$\mathcal{A} + \delta \;<\; D(PW||W_\star)\,. \tag{52}$$

Note that with such a choice we have in the left-hand side of (28)

$$\mathbb{E}\left[\exp\{-|r(Y^n) - \log A|^+\}\right] \le P[r(Y^n) < \mathcal{A} + \delta] + \exp\{-n\delta\}\,.$$

By the weak law of large numbers and (51)-(52) we have

$$P[r(Y^n) < \mathcal{A} + \delta] \;\to\; 0 \tag{53}$$

$$P[i(X^n; Y^n) > \gamma] \;\to\; 1\,. \tag{54}$$

Thus, the left-hand side of (45) converges to 1 and the constructed sequence of codes satisfies

$$\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W}, \hat{\nu} \neq \nu] \to 1\,. \tag{55}$$

■

C. Proof of Theorem 3

The key technical ingredient is the following generalization of a packing lemma [22, Lemma 2.5.1]:

Lemma 7: There exist a constant $K = K(|\mathcal{X}|, |\mathcal{Y}|) > 0$ and a positive integer $n = n_0(|\mathcal{X}|, |\mathcal{Y}|)$ such that for every $R > 0$, and every type P of sequences in $\mathcal{X}^n$ satisfying $H(P) > R$, there exist at least $M = \exp\{nR - K\log n\}$ distinct sequences $c_i \in \mathcal{X}^n$ of type P such that for every pair of stochastic matrices $V : \mathcal{X} \to \mathcal{Y}$, $\hat{V} : \mathcal{X} \to \mathcal{Y}$, every i and every $0 \le k < n$ we have

$$\left| T_V(c_i^{\star k}) \cap \bigcup_{j \neq i} T_{\hat{V}}(c_j) \right| \le \left|T_V(c_i^{\star k})\right| \exp\{-n|I(P, \hat{V}) - R|^+\} \tag{56}$$

provided that $n \ge n_0$.

Remark: In fact, there is nothing special about the transformations $c \mapsto c^{\star k}$. The lemma and the proof hold verbatim if $c_i^{\star k}$ is replaced by $f(c_i)$, and clause "every $0 \le k < n$" with "every

$f \in \mathcal{F}_n$", where $\mathcal{F}_n$ is an arbitrary collection of maps $f : \mathcal{X}^n \to \mathcal{X}^n$ of polynomial size: $|\mathcal{F}_n| = \exp\{O(\log n)\}$.

Proof: The argument is not much different from that of [22, Lemma 2.5.1] except that we need to account for the fact that $c_i^{\star k}$ may not have the type P. Below we demonstrate how the argument in [22] can be slightly modified to handle this generalization.

Let $M = \exp\{nR - K\log n\}$ with K to be specified later. It is clear that (56) is equivalent to the same statement with $\exp\{-n|\cdot|^+\}$ upper-bounded by $\exp\{-n(\cdot)\}$. For a given codebook $\mathcal{C}$ define

$$\lambda(i,k,V,\hat{V}) \stackrel{\triangle}{=} \frac{\left|T_V(c_i^{\star k}) \cap \bigcup_{j\neq i} T_{\hat{V}}(c_j)\right|}{\left|T_V(c_i^{\star k})\right|},$$

if the denominator is non-zero and we take $\lambda = 0$ otherwise. As in [22, Lemma 2.5.1] the proof will be complete if we can show for all $n \geq n_0(|\mathcal{X}|, |\mathcal{Y}|)$:

$$\mathbb{E}\left[\sum_{V,\hat{V},k} \lambda(1,k,V,\hat{V}) \exp\{n(I(P,\hat{V}) - R)\}\right] \leq \frac{1}{2}, \tag{57}$$

where the expectation is over a randomly generated codebook $\mathcal{C} = \{C_1, \ldots, C_M\}$ with C_j's being independent and uniformly distributed across the type T_P.

To compute $\mathbb{E}\left[\lambda(1,k,V,\hat{V})\right]$ assume that $T_V(c_1^{\star k}) \neq \emptyset$ and notice that then

$$\lambda(1,k,V,\hat{V}) = \mathbb{P}\left[U \in \bigcup_{j\neq 1} T_{\hat{V}}(C_j)\right],$$

where U is distributed uniformly across $T_V(C_1^{\star k})$. For this probability we have

$$\begin{aligned}
\mathbb{P}\left[U \in \bigcup_{j\neq 1} T_{\hat{V}}(C_j)\right] &\leq M\mathbb{P}[U \in T_{\hat{V}}(C_2)] && (58)\\
&= M\mathbb{P}[u_0 \in T_{\hat{V}}(C_2)] && (59)\\
&= M\frac{|\{x : u_0 \in T_{\hat{V}}(x)\}}{|T_P|} && (60)\\
&\leq M(n+1)^{|\mathcal{X}|}\exp\{-nI(P,\hat{V})\}, && (61)
\end{aligned}$$

where (58) is a union bound, in (59) we denoted by u_0 any element of $T_V(C_1^{\star k})$ and applied the permutation symmetry, (60) is because C_2 is uniform on T_P and (61) is by the same argument as in the proof of [22, Lemma 2.5.1].

Overall, regardless of $T_V(c_1^{\star k})$ being empty or not we have

$$\lambda(1,k,V,\hat{V})\exp\{n(I(P,\hat{V}) - nR)\} \leq (n+1)^{|\mathcal{X}|}\exp\{-K\log n\} \tag{62}$$

because of (61) and $M = \exp\{nR - K\log n\}$. Summing (62) over all k, V and $\hat{V}$ and applying the type-counting [22, Lemma 2.2] we see that for sufficiently large K the left-hand side in (57) converges to zero, hence there must exist n_0 such that (57) holds for all $n \geq n_0$. ■

We will also need the fact that over the finite space $Q \mapsto D(P||Q)$ is continuous (as an extended real-valued function):

Lemma 8: Let Q_1, Q_2 be distributions on a finite space $\mathcal{B}$ such that

$$q_{min} \stackrel{\triangle}{=} \min\{Q_1(y) : Q_1(y) > 0\} \tag{63}$$

$$\delta \stackrel{\triangle}{=} \max_{y\in\mathcal{B}} |Q_1(y) - Q_2(y)| < \frac{q_{min}}{2} \tag{64}$$

Then for every distribution P on $\mathcal{B}$ we have that $D(P||Q_1)$ and $D(P||Q_2)$ are finite or infinite simultaneously and in the former case

$$|D(P||Q_1) - D(P||Q_2)| \leq \frac{2\delta \log e}{q_{min}}\,.$$

Proof: Since $\delta < \frac{q_{min}}{2}$ it is clear that $Q_1 \sim Q_2$. Assuming $Q_1, Q_2 \ll P$ we get

$$D(P||Q_1) - D(P||Q_2) = \sum_{y\in\mathcal{B}} P(y)\log\frac{Q_1(y)}{Q_2(y)} \tag{65}$$

$$\leq \sum_{y\in\mathcal{B}} P(y)\frac{Q_1(y) - Q_2(y)}{Q_2(y)}\log e \tag{66}$$

$$\leq \frac{2\delta\log e}{q_{min}}\sum_{y\in\mathcal{B}} P(y)\,, \tag{67}$$

where (66) is by $\log x \leq (x-1)\log e$, and in (67) we applied

$$Q_1(y) - Q_2(y) \leq \delta\,, \tag{68}$$

$$Q_2(y) > \frac{q_{min}}{2}\,, \tag{69}$$

which follow by (64). Similar argument works for $D(P||Q_2) - D(P||Q_1)$. ■

Proof of Theorem 3: Let $P_n \to P$ be a sequence of n-types over $\mathcal{X}$ converging to P. The encoder for each n consists of the codebook constructed in Lemma 7 with composition P_n and

$$M_n = \exp\{nR - K\log n\}\,,$$

which clearly has asymptotic rate R.

We now describe the decoder, which operates in two phases:

1) Time slots $1,\ldots,n$ are used to estimate $W_\star$ by

$$\hat{W}_\star(b) \stackrel{\triangle}{=} \frac{1}{n}\sum_{j=1}^{k} 1\{Y_j = b\}\,, \qquad \forall b \in \mathcal{Y}\,.$$

2) From the time instant $t \geq 2n$ and on the decoder computes the conditional type of the block of last n letters with respect to every codeword $c_i, i = 1,\ldots,M$ and stops at the first moment when the conditional type enters the following set:

$$E_n = \{V : D(P_nV||\hat{W}_\star) > \mathcal{A} + n^{-c}, I(P_n,V) > R + n^{-c}\}\,,$$

where $0 < c < 1/4$ is an arbitrary fixed constant. Formally,

$$\hat{\nu} \stackrel{\triangle}{=} \inf\{t : \exists j : Y_t^{t+n-1} \in T_{E_n}(c_j)\}\,, \tag{70}$$

$$\hat{\mathcal{W}} \stackrel{\triangle}{=} \min\{j : Y_{\hat{\nu}}^{\hat{\nu}+n-1} \in T_{E_n}(c_j)\}\,. \tag{71}$$

This stopping rule may be seen as a simplified version of the one proposed in [14], properly adapted to the universal setting.

We now fix W satisfying (15)-(16) and use these encoders and decoders to construct a sequence of probability spaces generated using random transformation (W^n, A_n) according to (3) with

$$A_n = \exp\{\mathcal{A}n\}\,.$$

We upper-bound probability of error as follows:

$$\mathbb{P}[\{\hat{\nu} > \nu\} \cup \{\hat{\mathcal{W}} \neq \mathcal{W}\}] \tag{72}$$

$$\leq \mathbb{P}[\hat{\nu} > \nu] + \mathbb{P}[\hat{\mathcal{W}} \neq \mathcal{W}, \hat{\nu} \leq \nu - n] + \mathbb{P}[\hat{\mathcal{W}} \neq \mathcal{W}, \nu - n < \hat{\nu} \leq \nu] \tag{73}$$

$$\leq \mathbb{P}[\hat{\nu} > \nu] + \mathbb{P}[\hat{\nu} \leq \nu - n] + \mathbb{P}[\hat{\mathcal{W}} \neq \mathcal{W}, \nu - n < \hat{\nu} \leq \nu] \tag{74}$$

Next, we proceed to showing that each term in (74) tends to zero.

First, we show that with large probability the estimate $\hat{W}_\star$ is close to the actual $W_\star$. Denote the event

$$F \stackrel{\triangle}{=} \{\nu \leq 2n\} \cup \left\{||\hat{W}_\star - W_\star||_{TV} > n^{-1/4}\right\}\,,$$

where $||P - Q||_{TV} = \sum_{y\in\mathcal{Y}} |P(y) - Q(y)|$. Clearly, by the fact that

$$\mathbb{P}[\nu \leq 2n] = \frac{2n}{A_n}$$

Chebyshev bound and $\mathrm{Var}[W_\star(b)] = O(n^{-1})$ we have

$$\mathbb{P}[F] \to 0 \qquad n \to \infty\,. \tag{75}$$

Define an open neighborhood of W in the space of stochastic matrices as

$$B = \left\{V : I(P,V) > R + \frac{\epsilon_0}{2}, D(PV||W_\star) > \mathcal{A} + \frac{\epsilon_0}{2}\right\}\,,$$

where $\epsilon_0 > 0$ is chosen such that

$$I(P,W) > R + \epsilon_0, \quad D(PW||W_\star) > \mathcal{A} + \epsilon_0\,,$$

which is possible by (15)-(16). We now show that there exists an n_0 such that for all $n \geq n_0$ and every realization in F^c we have the inclusion

$$B \subset E_n\,. \tag{76}$$

Indeed, compactness of spaces of distributions on $\mathcal{X}$ and stochastic matrices $V : \mathcal{X} \to \mathcal{Y}$ implies uniform continuity of the map

$$(P,V) \mapsto I(P,V)\,.$$

Therefore, the sequence

$$\delta_n \stackrel{\triangle}{=} \max_V |I(P_n,V) - I(P,V)| \to 0 \tag{77}$$

is vanishing. Next, fix $V \in B$ and notice that by (77) we have

$$I(P_n,V) \geq I(P,V) - \delta_n > R + \frac{\epsilon_0}{2} - \delta_n\,.$$

Therefore, for all n sufficiently large we have

$$V \in B \implies I(P_n,V) > R + n^{-c}\,. \tag{78}$$

Assume that $V \in B$ is chosen such that $D(PV||W_\star) < \infty$. Consider the chain

$$\begin{aligned}
&|D(P_nV||W_\star) - D(PV||W_\star)| = && (79)\\
&\qquad \left|H(PV) - H(P_nV) + \sum_{y\in\mathcal{B}}(P_nV(y) - PV(y))\log\frac{1}{W_\star(y)}\right| && (80)\\
&\leq \; |H(PV) - H(P_nV)| + ||P_nV - PV||_{TV}\log\frac{1}{w_1} && (81)\\
&\leq \; \max_V |H(PV) - H(P_nV)| + ||P_n - P||_{TV}\log\frac{1}{w_1} && (82)\\
&= \; \delta_n'\,, && (83)
\end{aligned}$$

where in (81) we denoted

$$w_1 = \min\{W_\star(y) : W_\star(y) > 0\}\,,$$

and in (82) applied the data-processing for total variation. Notice now that by uniform continuity of entropy on the simplex of distributions on $\mathcal{B}$ we have

$$\delta_n' \to 0 \tag{84}$$

as $n \to \infty$ since $||P_n - P||_{TV} \to 0$. Therefore, for any realization in F^c and every V we have

$$\begin{aligned} D(PV||W_\star) &\leq D(P_nV||W_\star) + \delta_n' && (85)\\ &\leq D(P_nV||\hat{W}_\star) + \delta_n' + \frac{2\log e}{w_1}||P_n - P||_{TV}\,, && (86)\end{aligned}$$

where (85) is by (83) and (86) is by Lemma 8 with $Q_1 = W_\star$, $Q_2 = \hat{W}_\star$. Note that condition (64) is satisfied for all n sufficiently large since on F^c we have

$$||\hat{W}_\star - W_\star||_{TV} \leq n^{-1/4}\,. \tag{87}$$

Thus, by (86) and (84) we have for all sufficiently large n

$$V \in B, D(PV||W_\star) < \infty \implies D(P_nV||\hat{W}_\star) > \mathcal{A} + n^{-c} \tag{88}$$

In the case when $D(PV||W_\star) = \infty$ assume that n is large enough so that $P_n \sim P$. Then $D(P_nV||W_\star) = \infty$ and by Lemma 8 condition in the right-hand side of (88) is satisfied. Together (78), (88) and the previous argument imply (76).

Thus, for the first term in (74) we have

$$\begin{aligned} \mathbb{P}[\hat{\nu} > \nu] &\leq \mathbb{P}[F] + \mathbb{P}[\hat{\nu} > \nu, F^c] && (89)\\ &\leq \mathbb{P}[F] + \mathbb{P}[Y_\nu^{\nu+n-1} \notin T_{E_n}(c_{\mathcal{W}}), F^c] && (90)\\ &\leq \mathbb{P}[F] + \mathbb{P}[Y_\nu^{\nu+n-1} \notin T_{B}(c_{\mathcal{W}}), F^c] && (91)\\ &\leq \mathbb{P}[F] + 1 - \frac{1}{M}\sum_{j=1}^{M} W^n(T_B(c_j)|c_j) && (92)\\ &= o(1)\,, && (93)\end{aligned}$$

where (90) follows from the definition of $\hat{\nu}$ (note we used convention (11)), (91) is by (76), (92) is by (2), and (93) is by (75) and since B is a neighborhood of W and by [22, Lemma 2.12]

$$W^n\left(T_B(x^n)|x^n\right) \to 1\,, \tag{94}$$

for any sequence $x^n \in \mathcal{X}^n$.

Note that on F^c for any $V \in E_n$ we have for all n sufficiently large

$$\begin{aligned} D(P_nV||W_\star) &\geq D(P_nV||\hat{W}_\star) - \frac{2\log e}{w_1} n^{-1/4} && (95)\\ &> \mathcal{A} + n^{-c} - \frac{2\log e}{w_1} n^{-1/4} && (96)\\ &> \mathcal{A} + \frac{1}{2} n^{-c}\,, && (97)\end{aligned}$$

where (95) is by (87) and Lemma 8 assuming n is large so that (64) holds, (96) is by the assumption that $V \in E_n$ and (97) is by taking n large and recalling that $c < 1/4$.

Therefore, on F^c we have

$$y^n \in \cup_{j=1}^M T_{E_n}(c_j) \implies D(\hat{P}_{y^n}||W_\star) > \mathcal{A} + \frac{1}{2} n^{-c}\,,$$

where $\hat{P}_{y^n}$ denotes the $\mathcal{Y}$-type of the sequence $y^n = (y_1, \ldots, y_n)$, cf. [22, Definition 2.1]. Then we have

$$\begin{aligned} \mathbb{P}[Y_t^{t+n-1} \in \cup_{j=1}^M T_{E_n}(c_j) | t \leq \nu - n, F^c] &\leq \sum_{P_Y : D(P_Y||W_\star) > \mathcal{A} + \frac{1}{2} n^{-c}} \exp\{-nD(P_Y||W_\star)\} && (98)\\ &\leq (n+1)^{|\mathcal{Y}|} \exp\{-n\mathcal{A} - \frac{1}{2} n^{1-c}\}\,, && (99)\end{aligned}$$

where (98) follows from [22, Lemma 1.2.6] since for $t \leq \nu - n$ we have $Y_t^{t+n-1} \sim W_\star^n$, and (99) from counting of $\mathcal{Y}$-types.

Then, for the second term in (74) we have

$$\begin{aligned} \mathbb{P}[\hat{\nu} \leq \nu - n] &\leq \mathbb{P}[F] + \mathbb{P}[\exists t \leq \nu - n : Y_t^{t+n-1} \in \cup_{j=1}^M T_{E_n}(c_j), F^c] && (100)\\ &\leq \mathbb{P}[F] + A_n \mathbb{P}[Y_t^{t+n-1} \in \cup_{j=1}^M T_{E_n}(c_j) | F^c, t \leq \nu - n] && (101)\\ &\leq \mathbb{P}[F] + A_n (n+1)^{|\mathcal{Y}|} \exp\left\{-n\mathcal{A} - \frac{1}{2} n^{1-c}\right\} && (102)\\ &= o(1)\,, && (103)\end{aligned}$$

where (101) follows by the union bound, (102) is because of (99); and (103) is by (75) and $(n+1)^{|\mathcal{Y}|} \exp\{-n^{1-c}\} \to 0$.

Regarding the third term in (74) fix arbitrary $1 \le i \le M, 0 \le k < n$ and $\hat{V} : \mathcal{X} \to \mathcal{Y}$ and consider the following chain

$$\begin{aligned} W^n\left(\bigcup_{j\neq i} T_{\hat{V}}(c_j)\middle| c_i^{\star k}\right) &= \sum_V W^n\left(T_V(c_i^{\star k}) \cap \bigcup_{j\neq i} T_{\hat{V}}(c_j)\middle| c_i^{\star k}\right) && (104)\\ &\le \sum_V \exp\{-n|I(P,\hat{V}) - R|^+\} W^n(T_V(c_i^{\star k})|c_i^{\star k}) && (105)\\ &= \exp\{-n|I(P,\hat{V}) - R|^+\}\,, && (106)\end{aligned}$$

where (105) is by Lemma 7 and the fact that all strings in $T_V(c_i^{\star k})$ have the same probability. Thus, conditioning on E_n we get

$$\mathbb{P}[\hat{\mathcal{W}} \neq i | \hat{\nu} = \nu - k, E_n, \mathcal{W} = i] \le \sum_{\hat{V}\in E_n} \exp\{-n|I(P,\hat{V}) - R|^+\}\,, \qquad (107)$$

which follows from (106) and observing that when $\hat{\nu} = \nu - k$ the y^n block is effectively generated by the shift of a true codeword $c_i^{\star k}$ as defined in (27). Next, we obtain:

$$\begin{aligned} \mathbb{P}[\hat{\mathcal{W}} \neq \mathcal{W}, \nu - n < \hat{\nu} \le \nu] &\le \mathbb{P}[F] + \sum_{k=0}^{n-1} \mathbb{P}[\hat{\mathcal{W}} \neq \mathcal{W} | \hat{\nu} = \nu - k] && (108)\\ &\le \mathbb{P}[F] + n\mathbb{E}\left[\sum_{\hat{V}\in E_n} \exp\{-n|I(P_n,\hat{V}) - R|^+\}\right] && (109)\\ &\le \mathbb{P}[F] + (n+1)^{|\mathcal{X}|\,|\mathcal{Y}|+1} \exp\{-n^{1-c}\} && (110)\\ &= o(1)\,, && (111)\end{aligned}$$

where in (109) we average (107) over i and the realization of E_n, in (110) we used the type-counting [22, Lemma 2.2] and lower-bounded

$$I(P_n,\hat{V}) > R + n^{-c}$$

valid by the definition of E_n; and (111) is by (75) and $(n+1)^{|\mathcal{X}|\,|\mathcal{Y}|+1} \exp\{-n^{1-c}\} \to 0$.

Therefore, we have shown that in the upper-bound on probability of error (74) each term is $o(1)$ provided W is such that (15)-(16) hold. ■

IV. Non-asymptotic bound and channel dispersion

One of the important conclusions from the formula (13) is that the function $C(\mathcal{A})$ is constant on the interval $[0; \mathcal{A}_1]$, where $\mathcal{A}_1$ is given by (24). In other words, a certain level of asynchronism

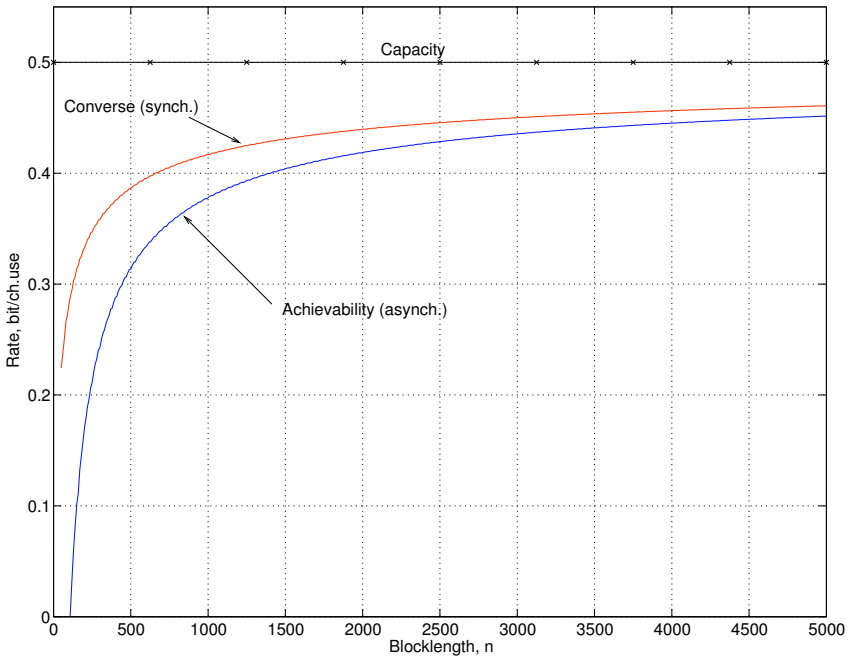

Fig. 2. BSC(0.11): Non-asymptotic performance of asynchronous codes compared with the upper (converse) bound for the synchronous channel. Probability of error $\epsilon = 10^{-3}$, asynchronism level $A_n = 2^{0.68n - 5.25\sqrt{n}}$.

(up to $\exp\{n\mathcal{A}_1\}$) is completely harmless to the capacity of the channel. This surprising result has also been noticed in [15] (the value of $\mathcal{A}_1$ is not known exactly for their model).

All the arguments so far were asymptotical and it is very natural to doubt whether such effect is actually possible for blocklengths of interest. To show that it does indeed happen for practical lengths we will employ the non-asymptotic achievability bound (Theorem 4). We will also demonstrate that for $\mathcal{A} \in [0, \mathcal{A}_1]$ neither the capacity nor the channel dispersion suffer any loss. First, however, we recall some of the results of [19].

Let $M^*(n, \epsilon)$ be the maximal cardinality of a codebook of blocklength n which can be (synchronously) decoded with block error probability no greater than ϵ over the DMC defined by (1). By Shannon's theorem asymptotically we have

$$\log M^*(n, \epsilon) \approx nC \tag{112}$$

It has been shown in [19] that a much tighter approximation can be obtained by defining an additional figure of merit referred to as the channel dispersion:

Definition 3: The dispersion V (measured in squared information units per channel use) of a channel with capacity C is equal to

$$V = \lim_{\epsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\frac{(nC-\log M^*(n,\epsilon))^2}{2\ln\frac{1}{\epsilon}}. \tag{113}$$

For example, the minimal blocklength required to achieve a given fraction η of capacity with a given error probability ϵ can be estimated as:[2]

$$n \gtrsim \left(\frac{Q^{-1}(\epsilon)}{1-\eta}\right)^2\frac{V}{C^2}. \tag{114}$$

The motivation for Definition 3 and estimate (114) is the following expansion for $n\to\infty$

$$\log M^*(n,\epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n). \tag{115}$$

As shown in [19] in the context of memoryless channels, (115) gives an excellent approximation for blocklengths and error probabilities of practical interest.

An interesting qualitative conclusion from Theorem 4 is the following:

Corollary 9: Consider a DMC W with (synchronous) capacity C and dispersion V. Then for every $0<\epsilon<1$ there exist capacity-dispersion optimal codes for the asynchronous DMC at asynchronism $A_n = 2^{n\mathcal{A}_1+o(n)}$. More precisely the number of messages M_n for such codes satisfies

$$\log M_n = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)\,, n\to\infty \tag{116}$$

and (4) holds.

Remark: As (115) demonstrates, it is not possible to improve the second term in expansion (116) even in the synchronous setting, see also [19, Theorem 48].

Proof: Apply Theorem 5 with the following choices

$$P_{X^n} = P^n\,, \tag{117}$$

$$A_n = \frac{1}{\sqrt{n}}\exp\{n\mathcal{A}_1 - n^{\frac{3}{4}}\sqrt{V_1}\} \tag{118}$$

$$\log M_n = nC - \sqrt{nV}Q^{-1}\left(\epsilon - \frac{B+4}{\sqrt{n}}\right) - \frac{3}{2}\log n\,, \tag{119}$$

$$\gamma_n = \log M_n + \frac{3}{2}\log n\,, \tag{120}$$

[2]As usual, $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-t^2/2}\,dt$.

where P is the capacity-dispersion achieving distribution,

$$V_1 = \mathrm{Var}[\log r(Y^n)]$$

and B is the Berry-Esseen constant for the sum of iid random variables comprising $i(X^n;Y^n)$ in (45); see [19, (259)]. With these choices we have for each of the terms in (45) for sufficiently large n:

$$\begin{aligned}
\mathbb{E}\left[\exp\{-|r(Y^n) - \log A_n|^+\}\right] &\leq \frac{2}{\sqrt{n}} && (121)\\
P[i(X^n;Y^n) > \gamma_n] &\geq 1 - \epsilon + \frac{4}{\sqrt{n}} && (122)\\
nM_n \exp\{-\gamma_n\} &\leq \frac{1}{\sqrt{n}} && (123)\\
(n-1)\exp\{-nE\} &\leq \frac{1}{\sqrt{n}}\,, && (124)
\end{aligned}$$

where (121) is by a simple estimate valid for all ν

$$\mathbb{E}\left[\exp\{-|r(Y^n) - \log A_n|^+\}\right] \leq P_{Y^n}[r(Y^n) < \nu] + A_n \exp\{-\nu\}$$

applied with $\nu = n\mathcal{A}_1 - n^{\frac{3}{4}}\sqrt{V_1}$ and first term estimated by Chebyshev; (122) is by an application of Berry-Esseen theorem, as in [19, (261)]; and (123)-(124) are obvious. Summing (121)-(124) we see that right-hand side of (45) is not smaller than $1-\epsilon$. Thus, by applying Taylor expansion to (119) we obtain (116). ■

Corollary 9 demonstrates that not only it is possible to communicate with rates close to capacity and still handle an exponential asynchronism (up to $2^{n\mathcal{A}_1}$), but in fact one can even do so using codes which are capacity-dispersion optimal.

Finally, in Fig. 2 we illustrate this last point numerically by computing the bound of Theorem 4 for the $BSC(\delta)$ and comparing it with the converse for the corresponding synchronous channel [19, Theorem 35]. For the purpose of this illustration we have chosen $\epsilon = 10^{-3}$, $\delta = 0.11$ and, somewhat arbitrarily,

$$\begin{aligned}
A_n &= \exp\left\{nD(P_Y^*||W_\star) + \sqrt{nV(P_Y^*||W_\star)}Q^{-1}(\tfrac{\epsilon}{4})\right\} && (125)\\
&\approx 2^{0.68n - 5.25\sqrt{n}}\,. && (126)
\end{aligned}$$

In particular, the plot shows that it is possible to construct asynchronous codes that do not lose much compared to the best possible synchronous codes in terms of rate, but which at the same

time are capable of tremendous tolerance to asynchronism. For example, already at $n = 500$ the decoder is able to find and error-correct the codeword inside a noisy binary string of unimaginable length $2^{221} \approx 10^{66}$.

We also remark that since the results of Theorem 4 and 5 only rely on the pairwise independence of the codewords in the ensemble, for the BSC when $M = 2^k$ we may use the ensemble corresponding to a random coset of a random linear code, that is the encoder function $f : \mathbb{F}_2^k \to \mathbb{F}_2^n$ is given by

$$f(b) = Gb + c_0, \qquad b \in \mathbb{F}_2^k \,,$$

where G is an $n \times k$ binary matrix with i.i.d. uniform entries and c_0 is a uniformly chosen element of $\mathbb{F}_2^k$. In this way, we conclude that for the BSC both the expansion (116) and the bound on Fig. 2 maybe achieved by a coset code. This naturally complements the results on coset codes in the classical models of back-to-back transmission [5], [6] and insertion-deletion [7], [9]; see Section I.

From the practical viewpoint, we thus expect that good modern error-correcting codes scrambled by a pseudo-random sequence will be a good solution to a problem of joint coding-synchronization.

ACKNOWLEDGMENT

We thank the anonymous reviewer for helpful comments regarding the prior work.

APPENDIX A
PROOF OF THE STRONG CONVERSE PART OF THEOREM 1

First, we introduce the performance of the optimal binary hypothesis test. Consider a $\mathcal{X}$-valued random variable X which can take probability measures P or Q. A randomized test between those two distributions is defined by a random transformation $P_{Z|X} : \mathcal{X} \mapsto \{0, 1\}$ where 0 indicates that the test chooses Q. The best performance achievable among those randomized tests is given by[3]

$$\beta_\alpha(P, Q) = \min \sum_{x \in \mathcal{X}} Q(x) P_{Z|X}(1|x) \,, \tag{127}$$

[3]We sometimes write summations over alphabets for simplicity of exposition; in fact, the definition holds for arbitrary measurable spaces.

where the minimum is over all probability distributions $P_{Z|X}$ satisfying

$$P_{Z|X} : \sum_{a\in\mathcal{X}} P(x)P_{Z|X}(1|x) \geq \alpha\,. \tag{128}$$

The minimum in (127) is guaranteed to be achieved by the Neyman-Pearson lemma. Thus, $\beta_\alpha(P,Q)$ gives the minimum probability of error under hypothesis Q if the probability of error under hypothesis P is not larger than $1-\alpha$. For more on the behavior of β_α see [25, Section 2.3] for example.

We proceed by noticing a pair of simple Lemmas.

Lemma 10: If $\mathcal{A}_\circ < \infty$ then there exists V_1 such that for any input x^n we have

$$\beta_\alpha(P_{Y^n|X^n=x^n}, W_\star^n) \geq \frac{\alpha}{2}\exp\left\{-nD(W||W_\star|\hat{P}_{x^n}) - \sqrt{\frac{2nV_1}{\alpha}}\right\}\,,$$

where $\hat{P}_{x^n}$ is the composition of x^n and

$$P_{Y^n|X^n=x^n}(\cdot) \stackrel{\triangle}{=} W^n(\cdot|x^n)\,,$$

with W^n defined in (1).

Proof: As in [19, Section IV.A] we define

$$V(W_x||W_\star) \stackrel{\triangle}{=} \sum_{y\in\mathcal{Y}} W(y|x)\log^2\frac{W(y|x)}{W_\star(y)} - D(W_x||Q)^2\,,$$

which is well-defined and finite whenever $\mathcal{A}_\circ < \infty$. Since $\mathcal{X}$ is finite we can set $V_1 = \max_{x\in\mathcal{X}} V(W_x||W_\star)$ and conclude by applying [19, Lemma 59]. ■

Lemma 11: Consider a DMC W. If $\mathcal{A}_\circ < \infty$ then there exists V_1 such that for any synchronous (n, M, ϵ) code (maximal probability of error) with codewords $\{c_i, i = 1,\dots M\}$ of constant composition P_0 we have

$$\beta_\alpha(P_{Y^n}, W_\star^n) \geq M\frac{\alpha}{4}\frac{\alpha-2\epsilon}{2-\alpha}\exp\left\{-nD(W||W_\star|P_0) - \sqrt{\frac{4nV_1}{\alpha-2\epsilon}}\right\} \tag{129}$$

provided that $\alpha > 2\epsilon$, where in (129) P_{Y^n} denotes the output distribution induced by the code:

$$P_{Y^n}[\cdot] = \frac{1}{M}\sum_{j=1}^{M} W^n(\cdot|c_i)\,.$$

Proof: Let $\delta = \alpha - \epsilon > 0$ and consider an arbitrary set A such that $P_{Y^n}[A] \geq \alpha$. Then by Chebyshev inequality at least

$$M' = \left\lceil \frac{\alpha}{2-\alpha} M \right\rceil \tag{130}$$

codewords $c_i \in \mathcal{X}^n$ satisfy

$$P_{Y^n|X^n=c_i}[A] \geq \frac{\alpha}{2}\,.$$

Without loss of generality, we assume they have indices $i = 1, \ldots, M'$. Let $D_i, i = 1, \ldots, M'$ denote the decoding regions corresponding to the given code. By the maximal probability of error requirement we have

$$P_{Y^n|X^n=c_i}[A \cap D_i] \geq \frac{\alpha}{2} - \epsilon\,.$$

Then

$$\begin{aligned} W_\star^n[A \cap D_i] &\geq \beta_{\frac{\alpha}{2}-\epsilon}(P_{Y^n|X^n=c_i}, W_\star^n) && (131)\\ &\geq \frac{\alpha - 2\epsilon}{4} \exp\left\{-nD(W||W_\star|P_0) + \sqrt{\frac{4nV_1}{\alpha - 2\epsilon}}\right\}, && (132)\end{aligned}$$

where (131) is by the definition of β_α and (132) by Lemma 10. Finally, we have

$$\begin{aligned} W_\star^n[A] &= \sum_{i=1}^{M} W_\star^n[A \cap D_i] && (133)\\ &\geq \sum_{i=1}^{M'} W_\star^n[A \cap D_i] && (134)\\ &\geq M\frac{\alpha}{2-\alpha}\frac{\alpha - 2\epsilon}{4} \exp\left\{-nD(W||W_\star|P_0) + \sqrt{\frac{4nV_1}{\alpha - 2\epsilon}}\right\}, && (135)\end{aligned}$$

where (133) is by disjointedness of D_i's and (135) is by (130) and (132). Extension from sets $A \subset \mathcal{Y}^n$ to random transformations $P_{Z|Y^n} : \mathcal{Y}^n \to \{0, 1\}$ is trivial and thus (129) follows by the definition of β_α. ■

Proof of the converse part of Theorem 1: We first consider the error definition as in (5). For the case $\mathcal{A}_\circ = \infty$ we can assume that a genie provides the value of ν to the decoder, in which case the problem becomes synchronous and the usual strong converse for the DMC applies. Thus, assume $\mathcal{A}_\circ < \infty$.

First, we assume $\epsilon < \frac{1}{3}$ and consider an (M_n, ϵ) code for a random transformation (W^n, A_n). By a standard argument, at the expense of a small increase in ϵ and a small decrease in M_n we can assume that (5) is replaced with

$$\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W}, \hat{\nu} \leq \nu | \mathcal{W} = j] \geq 1 - \epsilon\,, \qquad j = 1, \ldots, M_n\,. \tag{136}$$

We claim that such a code must be synchronously decodable over DMC W with maximal probability of error at most ϵ. Indeed, consider the following synchronous decoder: upon receiving y^n it generates ν uniform on $\{1,\dots,A_n\}$, puts the received y^n into slots $\nu\dots\nu+n-1$ and fills the rest of the slots with $W_\star$-generated noise. It then applies the given asynchronous decoder to the so-constructed element of $\mathcal{Y}^{A_n}$. Overall, by (136) the maximal probability of error must not exceed ϵ.

By another standard argument, e.g. [22, Chapter 10 and Theorem 10.6], there must exist a constant composition subcode with M_n' codewords such that for any $i=1,\dots,M_n'$ we have

$$\mathbb{P}[\hat{\mathcal{W}}=\mathcal{W}|\hat{\nu}=\nu,\mathcal{W}=i]\geq 1-\epsilon \tag{137}$$

and

$$\log M_n' \geq \log M_n - b_1\log n\,, \tag{138}$$

where $b_1>0$ is some (code-independent) constant. From now on we restrict attention to this subcode. It is well known that for some constant $b_3>0$ (depending only on W and ϵ)

$$\log M_n' \leq nI(P_n,W)+b_3\sqrt{n}\,, \tag{139}$$

where P_n is the composition of the code. Being interested in asymptotics, we are free to assume that

$$\frac{1}{M_n'}<1-3\epsilon\,. \tag{140}$$

We now apply the meta-converse principle [19, Section III.E], which consists of changing the channel and using (the complement of) the dominating error event as a binary hypothesis test between the two channels. Namely, in addition to the true channel $P_{Y^{A_n}|X^n,\nu}$ we consider an auxiliary channel

$$Q_{Y^{A_n}|X^n,\nu}=W_\star^{A_n}$$

which outputs $W_\star$-distributed noise in all of A_n symbols, regardless of X^n and ν. Obviously, under the Q-channel we have

$$\mathbb{Q}[\nu-n<\hat{\nu}\leq\nu]\leq\frac{n}{A_n} \tag{141}$$

by independence of $\hat{\nu}$ and ν, whereas under the P-channel we have

$$\begin{aligned}
\mathbb{P}[\hat{\nu} \le \nu, \hat{\mathcal{W}} = \mathcal{W}] &\le \mathbb{P}[\nu - n < \hat{\nu} \le \nu] + \mathbb{P}[\hat{\nu} \le \nu - n, \hat{\mathcal{W}} = \mathcal{W}] && (142)\\
&= \mathbb{P}[\nu - n < \hat{\nu} \le \nu] + \mathbb{P}[\hat{\nu} \le \nu - n]\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W} | \hat{\nu} \le \nu - n] && (143)\\
&\le \mathbb{P}[\nu - n < \hat{\nu} \le \nu] + \frac{1}{M_n'}\,, && (144)
\end{aligned}$$

where (144) is because $\mathbb{P}[\hat{\mathcal{W}} = \mathcal{W} | \hat{\nu} \le \nu - n] = \frac{1}{M}$ by conditional independence of $\hat{\mathcal{W}}$ and $\mathcal{W}$. Thus from (144) and (5) we have

$$\mathbb{P}[\nu - n < \hat{\nu} \le \nu] \ge 1 - \epsilon - \frac{1}{M_n'}\,. \tag{145}$$

Consider now the random variable $Z = 1\{\nu - n < \hat{\nu} \le \nu\}$. The kernel $P_{Z|Y^{A_n},\nu}$ acts from $\mathcal{Y}^{A_n} \times \{1, \dots, A_n - n + 1\}$ to $\{0, 1\}$ and constitutes a valid binary hypothesis test between $P_{Y^{A_n}\nu}$ and $Q_{Y^{A_n}\nu}$. Therefore,

$$\beta_{1-\epsilon'}(P_{Y^{A_n}\nu}, Q_{Y^{A_n}\nu}) \le \frac{n}{A_n}\,, \tag{146}$$

where we denoted for convenience

$$\epsilon' = \epsilon + \frac{1}{M_n'}\,.$$

On the other hand, for some $b_2 > 0$

$$\begin{aligned}
\beta_{1-\epsilon'}(P_{Y^{A_n}\nu}, Q_{Y^{A_n}\nu}) &= \beta_{1-\epsilon'}(P_{Y^{A_n}\nu}, W_\star^{A_n} \times P_\nu) && (147)\\
&= \beta_{1-\epsilon'}(P_{Y^{A_n}|\nu=1}, W_\star^{A_n}) && (148)\\
&= \beta_{1-\epsilon'}(P_{Y^n|\nu=1}, W_\star^{n}) && (149)\\
&\ge M_n' \exp\left\{-nD(W||W_\star|P_n) - b_2\sqrt{n}\right\}\,, && (150)
\end{aligned}$$

where (147) is by independence of Y^{A_n} and ν under Q, (148) is by [19, Lemma 29], (149) is because under $\nu = 1$ observations $Y_{n+1}^{A_n}$ are useless for discriminating the two hypothesis, and (150) is by Lemma 11 which is applicable because $1 - \epsilon' > 2\epsilon$ by (140).

Together (138), (139), (146) and (150) show that every code achieving probability of error ϵ over (W^n, A_n) must satisfy for some input type P_n:

$$\begin{aligned}
\log M_n &\le nD(W||W_\star|P_n) - \log A_n + b_2\sqrt{n} + b_1 \log n && (151)\\
\log M_n &\le nI(P_n, W) + b_3\sqrt{n} && (152)
\end{aligned}$$

The first implication from (151) is that there cannot be a sequence of $(n, 2, \epsilon)$ codes for asynchronism $A_n = \exp\{n\mathcal{A}\}$ with $\mathcal{A} > \max_x D(W_x||W_\star)$. Thus the synchronization thresholds $\mathcal{A}_{\circ,\epsilon}$ and $\mathcal{A}_\circ$ are given by (12).

Consider now a sequence of codes achieving $(R, \mathcal{A})$. The corresponding sequence of dominating types P_n must contain a subsequence converging to some P. Thus dividing by n and taking the limits in (151)-(152) we obtain that the pair $(R, \mathcal{A})$ must belong to

$$\bigcup_P \left\{ \begin{array}{rcl} R + \mathcal{A} & \leq & D(W||W_\star|P) \\ R & \leq & I(P, W) \end{array} \right\} . \tag{153}$$

Since for $\mathcal{A} \geq \mathcal{A}_\circ$ we already have shown $C(\mathcal{A}) = 0$ we can focus on $\mathcal{A} < \mathcal{A}_\circ$, for which we have:

$$\begin{aligned} C(\mathcal{A}) &= \max_P \min(D(W||W_\star|P) - \mathcal{A}, I(P, W) && (154)\\ &= \max_P \min(I(P, W) + D(PW||W_\star) - \mathcal{A}, I(P, W)) && (155)\\ &= \max_P I(P, W) - |\mathcal{A} - D(PW||W_\star)|^+ && (156)\\ &= \max_{P: D(PW||W_\star) \geq \mathcal{A}} I(P, W) \, , && (157) \end{aligned}$$

where (157) follows by noticing that the function under maximization in (156) is linear on the open set $\{D(PW||W_\star) < \mathcal{A}\}$ and equal to

$$D(W||W_\star|P) - \mathcal{A} \, . \tag{158}$$

Therefore, if the maximum in (156) were attained at a point P^* belonging to this open set, then we would have

$$D(W||W_\star|P^*) = \mathcal{A}_\circ \, . \tag{159}$$

Distribution P^* can not be concentrated on 1 atom, for otherwise $D(P^*W||W_\star) = \mathcal{A}_\circ > \mathcal{A}$, and therefore we can always modify P^* to increase the value of $D(P^*W||W_\star)$ until we have $D(P^*W||W_\star) = \mathcal{A}$, thereby showing that the maximum in (156) can be taken without loss of generality only over $\{D(PW||W_\star) \geq \mathcal{A}\}$.

To address the (practically uninteresting) case of $\epsilon \geq \frac{1}{3}$ we need to modify the proof as follows. To warrant applicability of Lemma 11 in (150) we needed to verify that

$$1 - \epsilon - \frac{1}{M'_n} \geq 2\epsilon_{max} \, , \tag{160}$$

where ϵ_{max} is the maximal probability (under synchronous decoding) of the considered code. To that end we perform an additional expurgation via [22, Corollary 2.1.9] to ensure $2\epsilon_{max} < 1-\epsilon$. Then for sufficiently large n (160) will hold and the rest of the argument is unchanged.

Finally, to address a generalization pointed out in the Remark after Theorem 2 and to handle a weaker condition (6) we need to replace the event $\{\nu - n < \hat{\nu}\}$ with

$$\left\{\nu - n < \hat{\nu} \le \nu + L_n, \frac{1}{n}\log\frac{1}{P_\nu(\nu)} > \mathcal{A} - \delta\right\},$$

where $\delta > 0$ is arbitrarily small. Then according to (14) we have

$$\mathbb{P}\left[\nu - n < \hat{\nu} \le \nu + L_n, \frac{1}{n}\log\frac{1}{P_\nu(\nu)} > \mathcal{A} - \delta\right] = \mathbb{P}\left[\nu - n < \hat{\nu} \le \nu + L_n\right] + o(1),$$

and hence the equivalent of the estimate (145) holds at the expense of arbitrary small enlargement of ϵ. In (141), on the other hand, we clearly have:

$$\mathbb{Q}\left[\nu - n < \hat{\nu} \le \nu + L_n, \frac{1}{n}\log\frac{1}{P_\nu(\nu)} > \mathcal{A} - \delta\right] \le (n + L_n)\exp\{-n\mathcal{A} + n\delta\}.$$

Continuing as above, we show (153) with $\mathcal{A}$ replaced by $\mathcal{A} - \delta$. Since the δ is arbitrary, the result follows. ■

REFERENCES

[1] R. H. Barker, "Group synchronization of binary digital systems," in *Communicatin Theory*, W. Jackson, Ed. New York: Academic-Butterworth, 1953.

[2] J. J. Stiffler, *Theory of Synchronous Communications*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[3] R. Scholtz, "Frame synchronization techniques," *IEEE Trans. Commun.*, vol. 28, no. 8, pp. 1204–1213, 1980.

[4] S. Golomb, B. Gordon, and L. R. Welch, "Comma-free codes," *Can. J. Math.*, vol. 10, pp. 202–209, 1958.

[5] J. J. Stiffler, "Comma-free error-correcting codes," *IEEE Trans. Inf. Theory*, vol. 11, no. 1, pp. 107–112, Jan. 1965.

[6] V. I. Levenshtein, "One method of constructing quasilinear codes providing synchronization in the presence of errors," *Prob. Peredachi Inform.*, vol. 7, no. 3, pp. 30–40, 1971.

[7] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[8] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Prob. Peredachi Inform.*, vol. 3, no. 4, pp. 11–26, Dec. 1967.

[9] R. Gallager, "Sequential decoding for binary channels with noise and synchronization errors," MIT, Lincoln Lab, Tech. Rep. AD0266879, Oct. 1961.

[10] G. Poltyrev, "Coding in an asynchronous multiple-access channel," *Prob. Peredachi Inform.*, vol. 19, no. 3, pp. 12–21, Sep. 1983.

[11] J. Hui and P. Humblet, "The capacity region of the totally asynchronous multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 31, no. 3, pp. 207–216, Mar. 1985.

[12] L. Farkas and T. Koi, "Capacity regions of discrete asynchronous multiple access channels," in *Proc. 2011 IEEE Int. Symp. Inf. Theory (ISIT)*, Aug. 2011, pp. 2273–2277.

[13] S. Verdú, "The capacity region of the symbol-asynchronous Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 35, no. 4, pp. 733 –751, Jul. 1989.

[14] A. Tchamkerten, V. Chandar, and G. W. Wornell, "Communication under strong asynchronism," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4508–4528, Oct. 2009.

[15] ——, "Asynchronous communication: Capacity bounds and suboptimality of training," *ArXiV*, 2011. [Online]. Available: http://arxiv.org/abs/1105.5639v1

[16] V. Chandar, A. Tchamkerten, and D. N. C. Tse, "Asynchronous capacity per unit cost," *IEEE Trans. Inf. Theory*, 2011, submitted. [Online]. Available: http://arxiv.org/abs/1007.4872v1

[17] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Stat.*, pp. 1897–1908, 1971.

[18] I. Nikiforov, "A generalized change detection problem," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 171–187, 1995.

[19] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[20] J. L. Massey, "Optimum frame synchronization," *IEEE Trans. Commun.*, vol. 20, no. 2, pp. 115–119, Apr. 1972.

[21] T. Cover, Y. Kim, and A. Sutivong, "Simultaneous communication of data and state," in *Proc. 2007 IEEE Int. Symp. Inf. Theory (ISIT)*. Nice, France: IEEE, 2007, pp. 916–920.

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

[23] V. Chandar, A. Tchamkerten, and G. Wornell, "Optimal sequential frame synchronization," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3725–3728, Aug. 2008.

[24] V. Chandar, A. Tchamkerten, and D. N. C. Tse, "Asynchronous capacity per unit cost," *arXiv*, Oct. 2012. [Online]. Available: http://arxiv.org/abs/1007.4872v2

[25] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 2010, available: http://www.princeton.edu/˜ypolyans.

[26] D. Wang, V. Chandar, S.-Y. Chung, and G. Wornell, "Error exponents in asynchronous communication," in *Proc. 2011 IEEE Int. Symp. Inf. Theory (ISIT)*. St. Petersburg, Russia: IEEE, 2011.

Yury Polyanskiy (S'08-M'10) received the M.S. degree (Hons.) in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia in 2005 and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2010. In 2000-2005 he lead development of the embedded software in the Department of Surface Oilfield Equipment, Borets Company LLC (Moscow). In 2011 Dr. Polyanskiy joined MIT as an Assistant Professor of Electrical Engineering and Computer Science (EECS), and a member of Laboratory of Information and Decision Systems.

His research interests include information theory, error-correcting codes, wireless communication and the theory of random processes. Over the years Dr. Polyanskiy won the 2011 Best Paper Award from IEEE Information Theory Society, the Best Student Paper Awards at the 2008 and 2010 IEEE International Symposiums on Information Theory (ISIT). His final year of graduate studies was supported by a Princeton University Honorific Dodds Fellowship (2009-2010). In 2012 Yury was selected to hold a Robert J. Shillman (1974) Career Development Professorship of EECS.