Lossless Source Coding in the Point-to-Point, Multiple Access, and Random Access Scenarios

Shuqing Chen, Michelle Effros, and Victoria Kostina
Dept. of Elec. Eng., California Institute of Technology, Pasadena, CA 91125 {*schen2, effros, vkostina*}*@caltech.edu*

***Abstract*—This paper treats point-to-point, multiple access and random access lossless source coding in the finite-blocklength regime. A random coding technique is developed, and its power in analyzing the third-order coding performance is demonstrated in all three scenarios. Results include a third-order-optimal characterization of the Slepian-Wolf rate region and a proof showing that for dependent sources, the independent encoders used by Slepian-Wolf codes can achieve the same third-order-optimal performance as a single joint encoder. The concept of random access source coding, which generalizes the multiple access scenario to allow for a subset of participating encoders that is unknown *a priori* to both the encoders and the decoder, is introduced. Contributions include a new definition of the probabilistic model for a random access-discrete multiple source, a general random access source coding scheme that employs a rateless code with sporadic feedback, and an analysis demonstrating via a random coding argument that there exists a deterministic code of the proposed structure that *simultaneously* achieves the third-order-optimal performance of Slepian-Wolf codes for all possible subsets of encoders.**

I. Introduction

This paper studies the finite-blocklength fundamental limits of fixed-length lossless source coding in three scenarios:

1) *Point-to-point*: A single source is compressed by a single encoder and decompressed by a single decoder.
2) *Multiple access*: Sources in a fixed set of sources are compressed by independent encoders and decompressed by a joint decoder.
3) *Random access*: Sources in an arbitrary subset of possible sources are compressed by independent encoders and decompressed by a joint decoder.

Following [1]–[4], we allow a non-vanishing error probability and study refined asymptotics of the achievable rates in encoding blocklength n.

In point-to-point almost-lossless source coding, non-asymptotic bounds and asymptotic expansions of the minimum achievable rate appear in [1], [3], [5]–[7]. In [3], Kontoyiannis and Verdú analyze the optimal code to give a *third-order*-optimal characterization of the minimum achievable rate $R^*(n,\epsilon)$ at blocklength n and error probability ϵ. For a finite-alphabet stationary memoryless source with single-letter distribution P_X, entropy $H(X)$, and varentropy $V(X)>0$,

$$R^*(n,\epsilon) \approx H(X)+\sqrt{\frac{V(X)}{n}}Q^{-1}(\epsilon)-\frac{\log n}{2n}, \qquad (1)$$

with any higher-order term bounded by $O(\frac{1}{n})$; here $Q^{-1}(\cdot)$ denotes the complementary Gaussian distribution function.

In multiple access lossless source coding, also known as Slepian-Wolf (SW) source coding [8], the object of interest is the set of achievable rate tuples, known as the rate region. The best prior asymptotic expansion of the SW rate region for a stationary memoryless multiple source is the *second-order*-optimal rate region, established independently in [9] and [10]. In [9], Tan and Kosut present a vector-form characterization of the SW rate region, which takes a form similar to the first two terms of (1). In this case, a quantity known as the entropy dispersion matrix plays a role similar to the varentropy $V(X)$. Their result suggests that the third-order term is bounded from above by $+O(\frac{\log n}{n})$ and from below by $-O(\frac{\log n}{n})$.

In this paper, we take a new approach to point-to-point almost-lossless source coding, combining random code design and maximum likelihood decoding to obtain a source coding counterpart to the random-coding union (RCU) bound from channel coding [2, Th. 16]. This new achievability bound (Theorem 1) yields a tight asymptotic expansion of $R^*(n,\epsilon)$ that achieves the first three terms of (1). The fact that our asymptotic expansion is achieved by a random code rather than the optimal code from [3] demonstrates that there is no loss (up to the third-order term) due to random code design, which implies that there are many good codes. Furthermore, our RCU bound can be generalized to source coding scenarios where the optimal code is not known; this is crucial since knowledge of the optimal code in the case of point-to-point almost-lossless source coding is quite exceptional.

While finding optimal SW codes is intractable in general, our derivation of the source coding RCU bound generalizes to SW source coding. The resulting achievability bound and a new converse based on composite hypothesis testing together yield the *third-order*-optimal rate region for SW source coding on a stationary memoryless multiple source (Theorem 2). Our result reveals a third-order term of $-\frac{\log n}{2n}$ that is independent of the number of encoders. This tightens the $+O(\frac{\log n}{n})$ third-order bound from [9], which grows linearly with the source alphabet size and exponentially with the number of encoders. Our third-order-optimal characterization implies that

This work is supported in part by the National Science Foundation under Grant CCF-1817241. The work of S. Chen is supported in part by the Oringer Fellowship Fund in Information Science and Technology.

S. Chen, M. Effros and V. Kostina are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA. E-mails: {*schen2, effros, vkostina*}*@caltech.edu.*

for dependent sources, the SW code's independent encoders suffer no loss up to the third-order performance relative to joint encoding with a point-to-point code.

The prior information theory literature studies multiple access source coding for scenarios where the number of encoders is fixed and known. In applications like sensor networks, the internet of things, and random access communication, however, the number of transmitters communicating with a given access point may be unknown or time-varying. The information theory of random access channel coding is investigated in papers such as [11]–[13]. Here, we introduce the notion of *random access (RA) source coding*, which extends multiple access source coding to scenarios where the set of active encoders is unknown *a priori*.

To begin our study, we first establish a probabilistic model for the object being compressed in RA source coding, here called the *random access-discrete multiple source* (RA-DMS). We then develop a robust coding scheme to achieve reliable compression of an arbitrary subset of the sources even when *a priori* knowledge of that subset is unavailable to the encoders and the decoder. Since the SW rate region varies with the source set, one might expect the encoders to vary their encoding strategy accordingly. In this case, however, the encoders do not know the source set, so we instead employ a rateless code. The encoders transmit their codewords symbol-by-symbol until the decoder informs them all to stop, with the decoder selecting a decoding time from a predetermined collection of potential decoding times based on the encoder activity pattern it observes in the network. Single-bit feedback from the decoder at each potential decoding time tells all encoders whether or not to continue transmitting.

We demonstrate (Theorem 3) that there exists a single deterministic code that *simultaneously* achieves, for all possible sets of active encoders, the third-order-optimal performance of SW codes. Since traditional random coding arguments do not demonstrate the existence of a single deterministic code that meets multiple independent constraints, prior code designs for multiple-constraint scenarios (see, for example, [14]) employ randomness shared between independent communicators. We here propose an alternative to that approach, deriving a refined random coding argument (Lemma 1) that demonstrates the existence of a single deterministic code that meets multiple constraints; this result eliminates the need for shared randomness in a variety of communication scenarios.

The paper is organized in the following way. In the rest of this section, we define notation. Section II, III, and IV are devoted to almost-lossless (point-to-point) source coding, SW (multiple access) source coding, and RA source coding, respectively. Except where noted, all source coding results presented here apply to both finite and countably infinite source alphabets. Further details appear in [15].

For any positive integer i, let $[i] \triangleq \{1, \dots, i\}$. We use upper case for random variables (e.g., X), lower case for scalar values (e.g., x), and both bold face and superscripts for vectors (e.g., $\mathbf{x} = x^n$, and $\mathbf{1} = (1, \dots, 1)$). Given vector $\mathbf{u} \in \mathbb{R}^d$ and set $\mathscr{S} \subset \mathbb{R}^d$, $\mathbf{u}+\mathscr{S}$ denotes the Minkowski sum of $\{\mathbf{u}\}$ and $\mathscr{S}$. Inequalities between two vectors of the same dimension indicate elementwise inequalities. Given a vector x^n and an ordered subset of its indices $\mathcal{T} \subseteq [n]$, we define $\mathbf{x}_\mathcal{T} \triangleq (x_i,\ i \in \mathcal{T}) \in \mathbb{R}^{|\mathcal{T}|}$. For any finite set $\mathcal{A}$, $\mathcal{P}(\mathcal{A})$ represents the power set of $\mathcal{A}$ excluding the empty set, giving $\mathcal{P}(\mathcal{A}) \triangleq \{\mathcal{T} : \mathcal{T} \subseteq \mathcal{A}\} \setminus \emptyset$. For functions $u(n)$ and $f(n)$, $u(n) = O(f(n))$ if there exist $c,\, n_0 \in \mathbb{R}^+$ such that $0 \le u(n) \le cf(n)$ for all $n > n_0$. For a multi-dimensional function $\mathbf{u} : \mathbb{N} \to \mathbb{R}^d$, $\mathbf{u}(n) = O(f(n))\mathbf{1}$ for some function $f(n)$ indicates that $u_i(n) = O(f(n))$ for all $i \in [d]$. All uses of 'log' and 'exp' employ an arbitrary common base, which determines the information unit.

Given an ordered set $\mathcal{T} \subset \mathbb{N}$, let $P_{\mathbf{X}_\mathcal{T}}$ be a distribution defined on countable alphabet $\mathcal{X}_\mathcal{T}$. For any $\mathcal{A}, \mathcal{B} \subseteq \mathcal{T}$ with $\mathcal{A} \cap \mathcal{B} = \emptyset$ and any $(\mathbf{x}_\mathcal{A}, \mathbf{x}_\mathcal{B}) \in \mathcal{X}_\mathcal{A} \times \mathcal{X}_\mathcal{B}$, the information and conditional information are defined as

$$\imath(\mathbf{x}_\mathcal{A}) \triangleq \log \frac{1}{P_{\mathbf{X}_\mathcal{A}}(\mathbf{x}_\mathcal{A})} \tag{2}$$

$$\imath(\mathbf{x}_\mathcal{A}|\mathbf{x}_\mathcal{B}) \triangleq \log \frac{1}{P_{\mathbf{X}_\mathcal{A}|\mathbf{X}_\mathcal{B}}(\mathbf{x}_\mathcal{A}|\mathbf{x}_\mathcal{B})}. \tag{3}$$

The corresponding (conditional) entropy, varentropy, and third centered moment of information are defined by, respectively,

$$H(\mathbf{X}_\mathcal{A}) \triangleq \mathbb{E}\left[\imath(\mathbf{X}_\mathcal{A})\right] \tag{4}$$

$$H(\mathbf{X}_\mathcal{A}|\mathbf{X}_\mathcal{B}) \triangleq \mathbb{E}\left[\imath(\mathbf{X}_\mathcal{A}|\mathbf{X}_\mathcal{B})\right] \tag{5}$$

$$V(\mathbf{X}_\mathcal{A}) \triangleq \mathrm{Var}\left[\imath(\mathbf{X}_\mathcal{A})\right] \tag{6}$$

$$V(\mathbf{X}_\mathcal{A}|\mathbf{X}_\mathcal{B}) \triangleq \mathrm{Var}\left[\imath(\mathbf{X}_\mathcal{A}|\mathbf{X}_\mathcal{B})\right] \tag{7}$$

$$T(\mathbf{X}_\mathcal{A}) \triangleq \mathbb{E}\left[|\imath(\mathbf{X}_\mathcal{A}) - H(\mathbf{X}_\mathcal{A})|^3\right] \tag{8}$$

$$T(\mathbf{X}_\mathcal{A}|\mathbf{X}_\mathcal{B}) \triangleq \mathbb{E}\left[|\imath(\mathbf{X}_\mathcal{A}|\mathbf{X}_\mathcal{B}) - H(\mathbf{X}_\mathcal{A}|\mathbf{X}_\mathcal{B})|^3\right]. \tag{9}$$

II. Almost-Lossless Source Coding

Definition 1 (Almost-Lossless Source Code)**.** *An (M, ϵ) code for a random variable X with alphabet $\mathcal{X}$ comprises an encoding function $\mathsf{f}\colon \mathcal{X} \to [M]$ and a decoding function $\mathsf{g}\colon [M] \to \mathcal{X}$ such that the error probability $\mathbb{P}\left[\mathsf{g}(\mathsf{f}(X)) \neq X\right] \le \epsilon$.*

Definition 2 (Block Almost-Lossless Source Code)**.** *A code for a random vector X^n defined on $\mathcal{X}^n$ is called an (n, M, ϵ) code.*

Definition 3 (Minimum Achievable Rate)**.** *The minimum code size and rate achievable at blocklength n and error probability ϵ are defined by, respectively,*

$$M^*(n, \epsilon) = \min\{M : \exists\, (n, M, \epsilon)\ \textit{code}\} \tag{10}$$

$$R^*(n, \epsilon) = \frac{1}{n} \log M^*(n, \epsilon). \tag{11}$$

Our new non-asymptotic achievability bound given in Theorem 1, stated next, is derived using i.i.d. uniform random codeword generation and maximum likelihood decoding.

Theorem 1 (RCU Bound)**.** *There exists an (M, ϵ) code for discrete random variable X such that*

$$\epsilon \le \mathbb{E}\left[\min\left\{1, \frac{1}{M}\mathbb{E}\left[\exp\left(\imath(\bar{X})\right)1\{\imath(\bar{X}) \le \imath(X)\}|X\right]\right\}\right], \tag{12}$$

where $P_{X\bar{X}}(a,b) = P_X(a)P_X(b)$ *for all* $a, b \in \mathcal{X}$.

Particularizing Theorem 1 to a stationary memoryless source with single-letter distribution P_X satisfying $V(X) > 0$ and $T(X) < \infty$ and invoking [2, Lemma 47] and the Berry-Esseen inequality gives an asymptotic achievabability bound on $R^*(n,\epsilon)$ that is identical to (1) in its first three terms.

III. Slepian-Wolf Source Coding

For notational brevity, we present our analysis on SW source coding for two encoders. All definitions and results generalize to scenarios with multiple encoders. (See [15].)

Definition 4 (SW Code)**.** *An* (M_1, M_2, ϵ) *SW code for a pair of random variables* (X_1, X_2) *defined on* $\mathcal{X}_1 \times \mathcal{X}_2$ *comprises encoding functions* $\mathsf{f}_1 \colon \mathcal{X}_1 \to [M_1]$ *and* $\mathsf{f}_2 \colon \mathcal{X}_2 \to [M_2]$ *and a decoding function* $\mathsf{g} \colon [M_1] \times [M_2] \to \mathcal{X}_1 \times \mathcal{X}_2$ *with error probability* $\mathbb{P}[\mathsf{g}(\mathsf{f}_1(X_1), \mathsf{f}_2(X_2)) \neq (X_1, X_2)] \leq \epsilon$.

Definition 5 (Block SW Code)**.** *A SW code for a pair of random vectors* (X_1^n, X_2^n) *defined on* $\mathcal{X}_1^n \times \mathcal{X}_2^n$ *is called an* (n, M_1, M_2, ϵ) *SW code.*

Definition 6 ((n,ϵ)-Rate Region)**.** *A rate pair* (R_1, R_2) *is* (n,ϵ)*-achievable if there exists an* (n, M_1, M_2, ϵ) *SW code with* $R_1 = \frac{1}{n}\log M_1$ *and* $R_2 = \frac{1}{n}\log M_2$*. The* (n,ϵ)*-rate region* $\mathscr{R}^*(n,\epsilon)$ *is the set of* (n,ϵ)*-achievable rate pairs.*

Let $\mathbf{Z}$ be a zero-mean Gaussian random vector in $\mathbb{R}^d$ with covariance matrix V. Define set

$$\mathscr{Q}_{\mathrm{inv}}(\mathsf{V}, \epsilon) \triangleq \{\mathbf{z} \in \mathbb{R}^d : \mathbb{P}[\mathbf{Z} \leq \mathbf{z}] \geq 1 - \epsilon\} \subset \mathbb{R}^d. \tag{13}$$

For any ordered set $\mathcal{T} \subset \mathbb{N}$, any distribution $P_{\mathbf{X}_\mathcal{T}}$ defined on $\mathcal{X}_\mathcal{T}$, and any $\mathbf{x}_\mathcal{T} \in \mathcal{X}_\mathcal{T}$, define $\left(2^{|\mathcal{T}|} - 1\right)$-dimensional vectors

$$\boldsymbol{\imath}_{\mathcal{P}(\mathcal{T})}(\mathbf{x}_\mathcal{T}) \triangleq \left(\imath\left(\mathbf{x}_\mathcal{A} | \mathbf{x}_{\mathcal{T}\setminus\mathcal{A}}\right),\, \mathcal{A} \in \mathcal{P}(\mathcal{T})\right) \tag{14}$$

$$\mathbf{H}_{\mathcal{P}(\mathcal{T})} \triangleq \mathbb{E}\left[\boldsymbol{\imath}_{\mathcal{P}(\mathcal{T})}(\mathbf{X}_\mathcal{T})\right], \tag{15}$$

and $\left(2^{|\mathcal{T}|} - 1\right) \times \left(2^{|\mathcal{T}|} - 1\right)$ matrix

$$\mathsf{V}_{\mathcal{P}(\mathcal{T})} \triangleq \mathrm{Cov}\left[\boldsymbol{\imath}_{\mathcal{P}(\mathcal{T})}(\mathbf{X}_\mathcal{T})\right]. \tag{16}$$

$\mathsf{V}_{\mathcal{P}(\mathcal{T})}$ is known as the entropy dispersion matrix (see [9, Def. 7]). For any vector $\mathbf{R}_\mathcal{T} \in \mathbb{R}^{|\mathcal{T}|}$, define the $\left(2^{|\mathcal{T}|} - 1\right)$-dimensional vector of its partial sums as

$$\overline{\mathbf{R}}_{\mathcal{P}(\mathcal{T})} \triangleq \left(\sum_{i \in \mathcal{A}} R_i,\, \mathcal{A} \in \mathcal{P}(\mathcal{T})\right). \tag{17}$$

Finally, define sets

$$\mathscr{R}^*_{\mathrm{in},\mathcal{T}}(n,\epsilon) \triangleq \Bigg\{\mathbf{R}_\mathcal{T} \in \mathbb{R}^{|\mathcal{T}|} : \overline{\mathbf{R}}_{\mathcal{P}(\mathcal{T})} \in \mathbf{H}_{\mathcal{P}(\mathcal{T})} + \frac{\mathscr{Q}_{\mathrm{inv}}\left(\mathsf{V}_{\mathcal{P}(\mathcal{T})}, \epsilon\right)}{\sqrt{n}} - \frac{\log n}{2n}\mathbf{1} + O\left(\frac{1}{n}\right)\mathbf{1}\Bigg\} \tag{18}$$

$$\mathscr{R}^*_{\mathrm{out},\mathcal{T}}(n,\epsilon) \triangleq \Bigg\{\mathbf{R}_\mathcal{T} \in \mathbb{R}^{|\mathcal{T}|} : \overline{\mathbf{R}}_{\mathcal{P}(\mathcal{T})} \in \mathbf{H}_{\mathcal{P}(\mathcal{T})} + \frac{\mathscr{Q}_{\mathrm{inv}}\left(\mathsf{V}_{\mathcal{P}(\mathcal{T})}, \epsilon\right)}{\sqrt{n}} - \frac{\log n}{2n}\mathbf{1} - O\left(\frac{1}{n}\right)\mathbf{1}\Bigg\}. \tag{19}$$

Our main result for SW source coding is presented below for two encoders.

Theorem 2 (Third-Order-Optimal SW Rate Region)**.** *Consider a pair of stationary memoryless sources with single-letter joint distribution* $P_{X_1X_2}$ *satisfying*

(a.1) $V(X_1|X_2) > 0, V(X_2|X_1) > 0, V(X_1, X_2) > 0$
(a.2) $T(X_1|X_2) < \infty, T(X_2|X_1) < \infty, T(X_1, X_2) < \infty$.

For any $0 < \epsilon < 1$*, the* (n,ϵ)*-rate region* $\mathscr{R}^*(n,\epsilon)$ *satisfies*

$$\mathscr{R}^*_{\mathrm{in},[2]}(n,\epsilon) \subseteq \mathscr{R}^*(n,\epsilon) \subseteq \mathscr{R}^*_{\mathrm{out},[2]}(n,\epsilon). \tag{20}$$

The proof of Theorem 2 applies an RCU bound to prove SW achievability and composite hypothesis testing with the asymptotics from [16] to establish a new converse.

When X_1 and X_2 are dependent, Theorem 2 implies that a SW code operating at rate point (R_1, R_2) on the boundary of $\mathscr{R}^*(n,\epsilon)$ achieves a sum rate that is equal (up to the third-order term) to the minimum achievable rate of a point-to-point code applied to vector source (X_1^n, X_2^n) provided that $R_1 < H(X_1)$ and $R_2 < H(X_2)$.

IV. Random Access Source Coding

Associate each encoder with a source from some fixed set of sources. In the RA source coding scenario, each encoder chooses whether to be *active* or not; only sources associated with the active encoders are compressed. We here establish the probabilistic model for the object being compressed in this scenario. Let $K < \infty$ be the maximal number of active encoders in the network and $\mathcal{T} \in \mathcal{P}([K])$ be an arbitrary ordered set.

Definition 7 (RA-DMS)**.** *An RA-DMS is a DMS where an unknown subset of sources is compressed. It is specified by joint distribution* $P_{\mathbf{X}_{[K]}}$ *such that when a subset of encoders indexed by* $\mathcal{T}$ *is active, the source distribution is the marginal*

$$P_{\mathbf{X}_\mathcal{T}}(\mathbf{x}_\mathcal{T}) = \sum_{\mathbf{x}_{[K]\setminus\mathcal{T}} \in \mathcal{X}_{[K]\setminus\mathcal{T}}} P_{\mathbf{X}_{[K]}}(\mathbf{x}_{[K]}),\ \forall\, \mathbf{x}_\mathcal{T} \in \mathcal{X}_\mathcal{T}. \tag{21}$$

We propose a communication scheme in the RA source coding scenario in which communication occurs in epochs. At the beginning of each epoch, each of the K encoders independently chooses whether to be active or not and retains its activity state until the end of the epoch. As a result, the set of active encoders $\mathcal{T}$ in a given epoch is fixed. In an epoch, each active encoder $i \in \mathcal{T}$ observes only its own source output X_i from a countable alphabet $\mathcal{X}_i$ and independently maps it to a codeword consisting of a sequence of *code symbols* drawn from code symbol alphabet $[Q_i]$. All of the $|\mathcal{T}|$ codewords are sent to the decoder symbol-by-symbol simultaneously. Since the set $\mathcal{T}$ of active encoders is unknown *a priori*, the encoder behavior cannot vary with it. The decoder, however, sees $\mathcal{T}$ and hence decides a time $m_\mathcal{T}$, called the *decoding blocklength*, at which it simultaneously decodes all the partial codewords it has received. The collection of potential decoding blocklengths $\mathcal{M} \triangleq (m_\mathcal{T} : \mathcal{T} \in \mathcal{P}([K]))$ is part of the code design and is known to all of the encoders and the decoder.

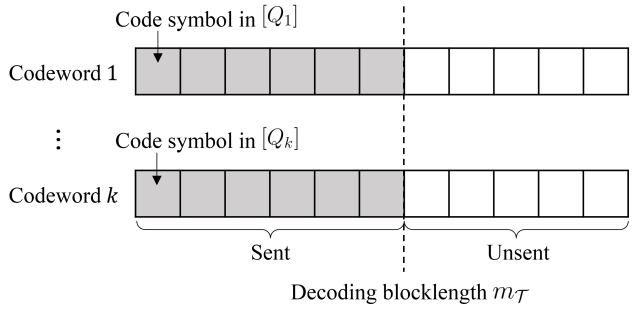

Fig. 1: Coding scheme in one epoch where the active encoder set $\mathcal{T} = [k]$.

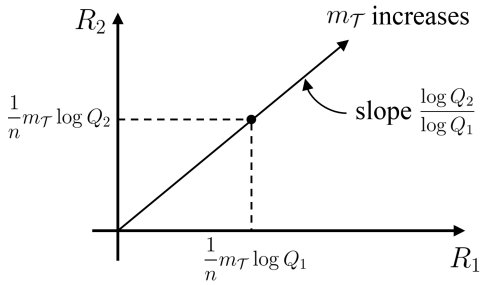

Fig. 2: The relationship between decoding blocklength $m_{\mathcal{T}}$, code symbol alphabet sizes (Q_1, Q_2), and source coding rate vector $\mathbf{R}_{\mathcal{T}}$, illustrated for $\mathcal{T} = [2]$.

Figure 1 illustrates our coding scheme in one epoch. At decoding blocklength $m_{\mathcal{T}}$, the decoder reconstructs the $|\mathcal{T}|$-dimensional source vector $\mathbf{X}_{\mathcal{T}}$ only from the first $m_{\mathcal{T}}$ code symbols sent from each active encoder and immediately tells those encoders to stop sending code symbols. In order to accomplish this termination of transmission, we let the decoder broadcast a single-bit acknowledgment (ACK) to all encoders at each time m in the set $\{m \in \mathcal{M} : m \leq m_{\mathcal{T}}\}$. For each $m < m_{\mathcal{T}}$, the decoder sends a "0" to indicate that it is not yet able to decode; in this case, the encoders keep sending code symbols. At time $m_{\mathcal{T}}$, the decoder sends a "1" to signal the end of one epoch and the start of the next. To avoid wasting time in an epoch with no active encoders, the decoder also sends an ACK at time $m_{\emptyset} = 1$ to signal whether ("0") or not ("1") there is any active encoder. As a result, when the active encoder set is $\mathcal{T}$, the encoders only need to tune in to receive ACKs at the predetermined times in the set $\left\{m \in \mathcal{M} : m \leq m^{\mathcal{T}}\right\}$ instead of listening to the feedback channel constantly. Given any possible set of active encoders, this scheme uses at most 2^K bits of feedback.

For the coding scheme described above, we define the following rateless code that can be employed in each epoch to accommodate any nonempty subset of active encoders. Define $\left(2^K - 1\right)$-dimensional vectors

$$\boldsymbol{\epsilon}_{\mathcal{P}([K])} \triangleq (\epsilon_{\mathcal{T}},\, \mathcal{T} \in \mathcal{P}([K])) \tag{22}$$

$$\mathbf{m}_{\mathcal{P}([K])} \triangleq (m_{\mathcal{T}},\, \mathcal{T} \in \mathcal{P}([K])) \tag{23}$$

and the maximal decoding blocklength

$$m_{\max} \triangleq \max\left\{m_{\mathcal{T}} : \mathcal{T} \in \mathcal{P}([K])\right\}. \tag{24}$$

Definition 8 (Random Access Source Code (RASC))**.** *An $\left(\mathbf{m}_{\mathcal{P}([K])}, \mathbf{Q}_{[K]}, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$ RASC for an RA-DMS with source alphabet $\mathcal{X}_{[K]}$ comprises a collection of encoding functions*

$$\mathsf{f}_i : \mathcal{X}_i \to [Q_i]^{m_{\max}},\, i \in [K], \tag{25}$$

where f_i is the encoding function employed by encoder i, and a collection of decoding functions

$$\mathsf{g}_{\mathcal{T}} : \prod_{i \in \mathcal{T}} [Q_i]^{m_{\mathcal{T}}} \to \mathcal{X}_{\mathcal{T}},\, \mathcal{T} \in \mathcal{P}([K]), \tag{26}$$

where $\mathsf{g}_{\mathcal{T}}$ is the decoding function used when the active encoder set is $\mathcal{T}$, such that for each $\mathcal{T} \in \mathcal{P}([K])$, source vector $\mathrm{X}_{\mathcal{T}}$ is decoded at time $m_{\mathcal{T}}$ with error probability $\mathbb{P}\left[\mathsf{g}_{\mathcal{T}}\left(\mathsf{f}_i(X_i)_{[m_{\mathcal{T}}]}, i \in \mathcal{T}\right) \neq \mathrm{X}_{\mathcal{T}}\right] \leq \epsilon_{\mathcal{T}}$. Here, $\mathsf{f}_i(x_i)_{[m_{\mathcal{T}}]}$ represents the first $m_{\mathcal{T}}$ code symbols of codeword $\mathsf{f}_i(x_i)$.

For each set of active encoders $\mathcal{T}$, the RASC reduces to a $\left((Q_i^{m_{\mathcal{T}}},\, i \in \mathcal{T}), \epsilon_{\mathcal{T}}\right)$ SW code (see Definition 4) with a finite number $|\{m \in \mathcal{M} : m \leq m_{\mathcal{T}}\}|$ of feedback bits. However, the RASC is *one* code that adopts a prefix structure (i.e., for each $x_i \in \mathcal{X}_i$, $\mathsf{f}_i(x_i)_{[m_{\mathcal{T}'}]}$ is a prefix of $\mathsf{f}_i(x_i)_{[m_{\mathcal{T}}]}$ if $m_{\mathcal{T}'} < m_{\mathcal{T}}$) and satisfies the error constraints for all $\mathcal{T} \in \mathcal{P}([K])$ simultaneously.

Definition 9 (Block RASC)**.** *An RASC for an RA-DMS with source alphabet $\mathcal{X}_{[K]}^n$ is called an $\left(n, \mathbf{m}_{\mathcal{P}([K])}, \mathbf{Q}_{[K]}, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$ RASC.*

Definition 10 (n-Valid Rate Collection)**.** *A collection of rate vectors $(\mathbf{R}_{\mathcal{T}})_{\mathcal{T} \in \mathcal{P}([K])}$, each indexed by its active encoder set $\mathcal{T}$, is n-valid if there exists a tuple $\left(\mathbf{m}_{\mathcal{P}([K])}, \mathbf{Q}_{[K]}\right)$ such that*

$$\mathbf{R}_{\mathcal{T}} = \frac{1}{n}\left(m_{\mathcal{T}} \log Q_i,\, i \in \mathcal{T}\right),\, \forall\, \mathcal{T} \in \mathcal{P}([K]). \tag{27}$$

Definition 10 reflects a key fact in the RASC design: while the decoding blocklength $m_{\mathcal{T}}$ can be chosen independently for each $\mathcal{T}$, the code symbol alphabet sizes $\mathbf{Q}_{[K]}$ are fixed and do not vary with the active encoder set. Thus, the rate vectors for different values of $\mathcal{T}$ are coupled as illustrated in Figure 2.

Definition 11 ($\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$-Rate Set)**.** *An n-valid rate collection $(\mathbf{R}_{\mathcal{T}})_{\mathcal{T} \in \mathcal{P}([K])}$ is $\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$-achievable if there exists an $\left(n, \mathbf{m}_{\mathcal{P}([K])}, \mathbf{Q}_{[K]}, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$ RASC. The $\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$-rate set $\mathcal{R}^*\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$ is the set of $\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$-achievable rate collections.*

Theorem 3 presents our third-order-optimal characterization of the $\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$-rate set. Define sets

$$\mathcal{R}_{\text{in}}^*\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right) \triangleq \Big\{n\text{-valid}\,(\mathbf{R}_{\mathcal{T}})_{\mathcal{T} \in \mathcal{P}([K])} : \mathbf{R}_{\mathcal{T}} \in \mathscr{R}_{\text{in},\mathcal{T}}^*(n, \epsilon_{\mathcal{T}}),\, \forall\, \mathcal{T} \in \mathcal{P}([K])\Big\} \tag{28}$$

$$\mathcal{R}_{\text{out}}^*\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right) \triangleq \Big\{n\text{-valid}\,(\mathbf{R}_{\mathcal{T}})_{\mathcal{T} \in \mathcal{P}([K])} : \mathbf{R}_{\mathcal{T}} \in \mathscr{R}_{\text{out},\mathcal{T}}^*(n, \epsilon_{\mathcal{T}}),\, \forall\, \mathcal{T} \in \mathcal{P}([K])\Big\}, \tag{29}$$

where $\mathscr{R}_{\text{in},\mathcal{T}}^*(n, \epsilon)$ and $\mathscr{R}_{\text{out},\mathcal{T}}^*(n, \epsilon)$ are the third-order bounding SW sets for source distribution $P_{\mathbf{X}_{\mathcal{T}}}$ (see (18) and (19)).

Theorem 3 (Third-Order-Optimal Performance of RASC)**.** *For any $K < \infty$, consider a stationary memoryless RA-DMS specified by single-letter joint distribution $P_{\mathbf{X}_{[K]}}$ that satisfies*

(b.1) $V(\mathbf{X}_{\mathcal{A}}|\mathbf{X}_{\mathcal{T}\setminus\mathcal{A}}) > 0$, $\forall\, \mathcal{A} \subseteq \mathcal{T} \subseteq [K]$, $\mathcal{A}, \mathcal{T} \neq \emptyset$
(b.2) $T(\mathbf{X}_{\mathcal{A}}|\mathbf{X}_{\mathcal{T}\setminus\mathcal{A}}) < \infty$, $\forall\, \mathcal{A} \subseteq \mathcal{T} \subseteq [K]$, $\mathcal{A}, \mathcal{T} \neq \emptyset$.

For any $\mathbf{0} < \boldsymbol{\epsilon}_{\mathcal{P}([K])} < \mathbf{1}$, the $\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$-rate set satisfies

$$\mathcal{R}^*_{\text{in}}\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right) \subseteq \mathcal{R}^*\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right) \subseteq \mathcal{R}^*_{\text{out}}\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right). \tag{30}$$

It follows from Theorem 3 that given any $\mathbf{Q}_{[K]}$, we can always find an $\mathbf{m}_{\mathcal{P}([K])}$ that yields a collection of rate vectors where each rate vector $\mathbf{R}_{\mathcal{T}}$ for $\mathcal{T} \in \mathcal{P}([K])$ gives a point on the boundary of the third-order-optimal SW rate region corresponding to $\mathcal{T}$. Each $\mathbf{Q}_{[K]}$ determines *one* collection of such boundary points. Therefore, on a class of stationary memoryless RA-DMSs that satisfy (b.1)-(b.2), our rateless coding scheme, which is agnostic to the set of active encoders *a priori*, is able to perform as well (up to the third order term) as a collection of SW codes with the same code symbol alphabets, where each SW code is optimally designed for a known active encoder set $\mathcal{T} \in \mathcal{P}([K])$.

An RASC with $Q_i = Q$ for all $i \in [K]$ operates at the symmetrical rate point for every $\mathcal{T} \in \mathcal{P}([K])$. If the source distribution $P_{\mathbf{X}_{[K]}}$ is permutation-invariant and satisfies $\mathbb{P}\big[\bigcup_{i,j\in[K],\, i\neq j}\{X_i = X_j\}\big] < 1$, we can significantly reduce the complexity of the code design by employing identical encoding for all encoders and identity-blind decoding (see [15]).

A. Converse Proof of Theorem 3

When analyzing the converse of the RASC, we relax the constraints by allowing prior knowledge of the active encoder set and exactly 2^K bits of feedback. We show that any *constant* bits of feedback does not change the SW rate region in the first three terms. Thus, we obtain the converse for the RASC from the converse for the SW code for each $\mathcal{T} \in \mathcal{P}([K])$.

B. Achievability Proof of Theorem 3

We first analyze rate collections that are $\left(n, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$-achievable when the encoders and decoder share common randomness used to generate a random code. Since proving the existence of a random code ensemble with expected error probabilities satisfying all error constraints does not guarantee the existence of a *single* deterministic code satisfying those constraints simultaneously, we require a new approach.

Lemma 1 below bounds the probability (with respect to a random code ensemble) that the error probability of a randomly chosen code exceeds a certain threshold.

Lemma 1. *Let $\mathcal{C}$ be any class of codes. For any code $\mathsf{c} \in \mathcal{C}$, let $P_e(\mathsf{c})$ denote the error probability associated with code c, and let*

$$\epsilon^*(\mathcal{C}) \triangleq \min_{\mathsf{c}\in\mathcal{C}} P_e(\mathsf{c}) \tag{31}$$

denote the error probability of the best code in $\mathcal{C}$. Then any random code ensemble C defined over $\mathcal{C}$ satisfies

$$\mathbb{P}\left[P_e(\mathsf{C}) > \epsilon\right] \leq \frac{\mathbb{E}\left[P_e(\mathsf{C})\right] - \epsilon^*(\mathcal{C})}{\epsilon - \epsilon^*(\mathcal{C})},\ \forall\, \epsilon > \epsilon^*(\mathcal{C}). \tag{32}$$

Let $P_{e,\mathcal{T}}(\mathsf{c})$ denote the error probability of RASC code c provided that the active encoder set is $\mathcal{T}$, for each $\mathcal{T} \in \mathcal{P}([K])$. When applying Lemma 1, we use our bound on each $\mathbb{E}[P_{e,\mathcal{T}}(\mathsf{C})]$ and the minimal error probability of the corresponding SW code at a given choice of $\mathbf{m}_{\mathcal{P}([K])}$ and $\mathbf{Q}_{[K]}$ to evaluate the probability

$$\mathbb{P}\left[\bigcup\nolimits_{\mathcal{T}\in\mathcal{P}([K])}\left\{P_{e,\mathcal{T}}(\mathsf{C}) > \epsilon_{\mathcal{T}}\right\}\right] \tag{33}$$

that a randomly drawn RASC C has error probability $P_{e,\mathcal{T}}(\mathsf{C})$ greater than $\epsilon_{\mathcal{T}}$ for some $\mathcal{T}$. We show that for any $\mathbf{Q}_{[K]}$, with a choice of $\mathbf{m}_{\mathcal{P}([K])}$ such that

$$m_{\mathcal{T}} = \min\left\{m : \frac{1}{n}(m \log Q_i, i \in \mathcal{T}) \in \mathscr{R}^*_{\text{in},\mathcal{T}}(n, \epsilon_{\mathcal{T}})\right\} \tag{34}$$

for each $\mathcal{T} \in \mathcal{P}([K])$, we can bound the probability of each event in the union in (33) from above by 2^{-K}, which makes (33) strictly less than 1, implying the existence of a deterministic $\left(n, \mathbf{m}_{\mathcal{P}([K])}, \mathbf{Q}_{[K]}, \boldsymbol{\epsilon}_{\mathcal{P}([K])}\right)$ RASC.

References

[1] V. Strassen, "Asymptotische abschäzungen in Shannons informationstheorie," in *Proc. Trans. Third Prague Conf. Inf. Theory, Statist., Decision Funct., Random Process.*, 1964, p. 689–723.

[2] 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.

[3] I. Kontoyiannis and S. Verdú, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.

[4] V. Kostina, Y. Polyanskiy, and S. Verdú, "Variable-length compression allowing errors," *IEEE Trans. Inf. Theory*, vol. 61, no. 8, pp. 4316–4330, Aug. 2015.

[5] A. A. Yushkevich, "On limit theorems connected with the concept of entropy of markov chains," *Uspekhi Mat. Nauk*, vol. 8, no. 5(57), p. 177–180, 1953.

[6] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer-Verlag Berlin Heidelberg, 2003.

[7] M. Hayashi, "Second-order asymptotics in fixed-length source coding and intrinsic randomness," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4619–4637, Aug. 2008.

[8] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[9] V. Y. F. Tan and O. Kosut, "On the dispersions of three network information theory problems," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 881–903, Feb. 2014.

[10] R. Nomura and T. S. Han, "Second-order Slepian-Wolf coding theorems for non-mixed and mixed sources," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5553–5572, Sep. 2014.

[11] P. Minero, M. Franceschetti, and D. N. C. Tse, "Random access: An information-theoretic perspective," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 909–930, Feb. 2012.

[12] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2523–2527.

[13] M. Effros, V. Kostina, and R. C. Yavas, "Random access channel coding in the finite blocklength regime," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1261–1265.

[14] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[15] S. Chen, M. Effros, and V. Kostina, "Lossless source coding in the point-to-point, multiple access, and random access scenarios," *ArXiv preprint*, 2019.

[16] Y. Huang and P. Moulin, "Strong large deviations for composite hypothesis testing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2014, pp. 556–560.