Lossless Coding for Distributed Streaming Sources

Stark C. Draper, *Member, IEEE*, Cheng Chang, and Anant Sahai, *Member, IEEE*

***Abstract*— Distributed source coding is traditionally viewed in a block coding context wherein all source symbols are known in advance by the encoders. However, many modern applications to which distributed source coding ideas are applied, are better modeled as having streaming data. In a streaming setting, source symbol pairs are revealed to separate encoders in real time and need to be reconstructed at the decoder with subject to some tolerable end-to-end delay. In this paper, a causal sequential random binning encoder is introduced and paired with maximum likelihood (ML) and universal decoders. The latter uses a novel weighted empirical suffix entropy decoding rule. We derive a lower bounds on the error exponent with delay for each decoder. We also provide upper bounds for the special case of streaming with decoder side information and discuss when upper and lower bounds match. We show that both ML and universal decoders achieve the same (positive) error exponents for all rate pairs inside the Slepian–Wolf achievable rate region. The dominant error events in streaming are different from those in block-coding and result in different exponents. Because the sequential random binning scheme is also universal over delays, the resulting code eventually reconstructs every source symbol correctly with probability one.**

***Index Terms*— Distributed source coding, lossless source coding, Slepian–Wolf coding, streaming data, universal decoding.**

I. Introduction

DISTRIBUTED source coding, pioneered in its lossless form by Slepian and Wolf [27] and in its lossy form by Wyner and Ziv [31], has found use in many applications; see [6] for a survey. While the standard distributed source coding paradigm is block-oriented, the data characteristics particular to certain applications of interest, e.g., video codecs [9], [14], [19], [20], are streaming in nature. Rather than being fully realized in advance, the data is realized and encoded in real time. This motivates the investigation of extensions of distributed source coding to streaming sources which, in its lossless variant, is the focus of this paper.

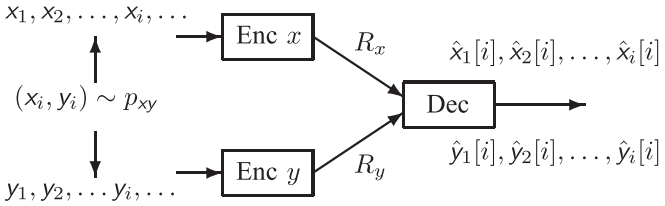

Fig. 1. Streaming distributed source coding. At time i a source pair (x_i, y_i) is received at the encoder, R_x and R_y bits are sent by the respective encoders to the joint decoder, and an estimate of the jth pair $(\hat{x}_j[i], \hat{y}_j[i])$, for all $j \leq i$, is made. The delay on this estimate is $\Delta = i - j$. We bound individual and joint error probabilities as a function of source statistics, R_x, R_y, and Δ.

The streaming version of the Slepian-Wolf problem studied in this paper is illustrated in Fig. 1. The sources are modeled as being embedded in time, integrating the idea that all physically realizable encoders/decoders must obey some form of causality. Encoders do not have access to the entire source realization in advance, rather source symbols continue to arrive at the encoder during the course of transmission.

Within the model of Fig. 1 we desire a probability of error that goes to zero for every source symbol, but at the cost of variable delay. In other words, consider the jth pair (x_j, y_j). At any time $i \geq j$ we want the probability that the estimate we can make at that time $(\hat{x}_j[i], \hat{y}_j[i])$ is not correct to drop exponentially in the delay $\Delta = i - j$. Achieving such "anytime" reliability turns out to be key in a number of distributed control and coordination problems (see, e.g., [21], [23]). While those earlier works on anytime reliability focused on channel coding, herein we ask analogous questions of distributed source coding.

In this paper, we formally define a streaming Slepian-Wolf code, and develop coding strategies both for situations when source statistics are known and when they are not. The new tool we introduce is a sequential binning argument that parallels the tree-coding arguments used to study convolutional codes. We characterize the performance of the streaming schemes through an error-exponent analysis and demonstrate the same exponents can be achieved regardless of whether the system is informed of the source statistics (in which case we use maximum-likelihood (ML) decoding) or not (in which case we use universal decoding). The universal decoder we design for the streaming problem is somewhat different from those familiar from the block coding literature, as are the nature of the error exponents in both the universal and ML cases. The end results is that, essentially, every source symbol can *eventually* be recovered correctly with probability one. In

Manuscript received August 5, 2010; revised September 21, 2013; accepted November 1, 2013. Date of publication December 6, 2013; date of current version February 12, 2014. This work was supported in part by the National Science Foundation (NSF) CAREER under Grant CCF-0844539, in part by the NSF ITR Grant CNS-0326503, and in part by NSF Grant CCF-0729122. This paper was presented at the 2005 IEEE International Symposium on Information Theory [8].

S. C. Draper is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: stark.draper@utoronto.ca).

C. Chang was with the Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA 94720 USA. He is now with D. E. Shaw, New York, NY 10036 USA (e-mail: cchang@eecs.berkeley.edu).

A. Sahai is with the Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA 94720 USA (e-mail: sahai@eecs.berkeley.edu).

Communicated by E.-H. Yang, Associate Editor for Source Coding.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2294368

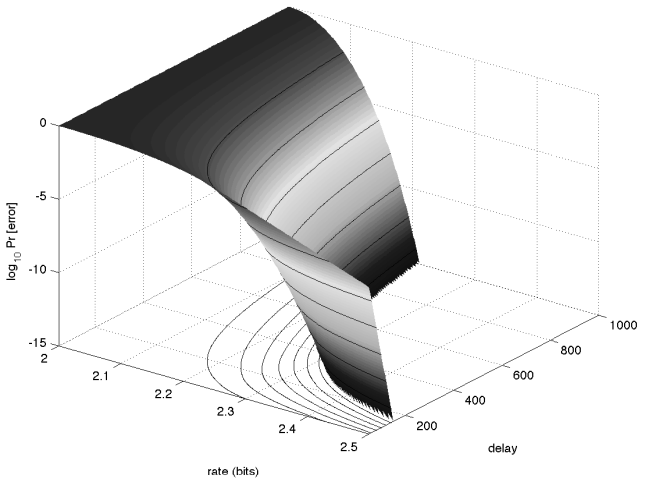

Fig. 2. The achievable tradeoff between rate, delay, and probability of error for an i.i.d. source that has a 50% chance of emitting a 0 and a 12.5% chance of emitting a $1, 2, 3$, or a 4. The entropy is 2 bits. The surface represents what the bounds in this paper achieve.

particular, at any time the decoder can make a causal estimate of *any* specific symbol. The decoder can continue to refine these estimates over time. The bound on erroneous estimation decays exponentially in the delay, Δ. The choice of acceptable delay is up to the user, based on application requirements.

From an engineering perspective, four desirable qualities of our scheme would be: (i) low-rate transmission, (ii) small end-to-end latency, (iii) low probability of error, and (iv) low implementational complexity. As is often the case in information theoretic investigations, we will not consider implementation complexity. The theory we develop does tell us about the tradeoffs among the first three of these qualities. In Fig. 2 we illustrate the tradeoff between rate, latency, and error probability that is revealed by our analysis for a bursty discrete memoryless source. For simplicity we plot results for a point-to-point streaming system which can be understood as the system illustrated in Fig. 1 in the special case where y is independent of x. In the example, with probability 0.5 the realization of each i.i.d. source symbols x_i is 0 and with probability 0.5 the realization is uniformly distributed across $\{1, 2, 3, 4\}$. The entropy of this source is 2 bits. The surface plotted in the figure depicts the upper bound on achievable error probability derived in this paper as a function of rate and delay.

A. Relation to Prior Work

The system depicted in Fig. 1 is related to models of delay-constrained source coding studied previously. Perhaps the most closely related work is that by Weissman and El Gamal [29]. In [29] the authors consider a variant of the source coding with side-information problem wherein the "side-information" sequence y_i is revealed directly to the decoder (rather than through a rate-limited channel as in Fig. 1). In [29] the encoder observes the full length-n source realization non-causally. The decoder, however, must operate causally (or with some look-ahead), estimating source symbol x_i based on the message from the encoder and the side-information sequence up to some l steps in the future: $y_1, \ldots, y_{i+l}$. The authors find the somewhat pessimistic results that any finite look-ahead of l is useless in the sense that the encoding rate R_x must satisfy $R_x > H(x)$ (rather than $R_x > H(x|y)$) to ensure that $\lim_{n\to\infty} \Pr(x^n \neq \hat{x}^n) = 0$. This may seem at first to be at odds with the results of this paper. We show that we can attain an exponential decay in error, i.e., $\Pr(\hat{x}_i[i+l] \neq x_i) \leq 2^{-lE(R_x)}$ with positive exponent $E(R_x) > 0$ as long as $R_x > H(x|y)$. The resolution is that our reliability is exponential in *delay* and not in the absolute position of the symbol to be estimated in the sequence so, indeed, $\lim_{i\to\infty} \Pr(\hat{x}_i[i+l] \neq x_i) \leq 2^{-lE(R_x)} \neq 0$.

Another relevant set of work concerns variable-length Slepian-Wolf coding. In this setting either codeword lengths or the number of bits transmitted are a function of the realized source sequence. Our setting is slightly different as our encoders operate at fixed rates and it is the decision time that can be variable. While one use of variable-length coding in classic source-coding is to attain zero-error compression (e.g., by using Huffman codes) variable-length coding does *not*, in general, enable zero-error Slepian-Wolf coding at rates close to the conditional entropy. This result can be inferred from the interactive data compression setting of [13] where it is shown that, in general, a zero-error variant of the source-coding with side-information problem is possible only if $R_x \geq H(x)$. Applying the result twice reveals that $R_x + R_y \geq H(x) + H(y)$ to get zero-error for the Slepian-Wolf problem of Fig. 1. One should note that, just as is the case for zero-error channel coding, when certain symbol pairs are known to have zero probability, there are special cases where zero-error Slepian-Wolf coding is possible [17]. But, while not getting to zero error, variable-length coding can sometimes help in a second-order sense. A Slepian-Wolf code with codewords of different lengths is used in [15] to reduce the redundancy of the code, i.e., the rate above the conditional entropy used at finite n. The extra usefulness of variable over fixed-length coding depends somewhat naturally on a combination of the non-uniformity of the x-source and the conditional deviation of the output from its marginal given each source observation. These two effects interact, see Remark 10 of [15] and some remarks following Theorem 5 in Section III. Finally, we note that if interaction is allowed between the encoder and decoder, then variable-length approaches have been studied that adapt the encoding rates in an on-line manner to deal with, say, unknown statistics. See, e.g., [7], [10], [26], [32].

Finally, we note that recently constructions of causal encoders of the type considered in this paper have been considered. Since the analysis herein depends only on pairwise error probability, linear codes suffice. Causal encoding and a linear structure means that the parity-check matrix of the code must have a lower-diagonal design. Such a design in the somewhat different context of interactive source coding with decoder side information is considered in [18]. More closely related is the discussion of linear anytime codes considered recently in [28]. By constraining the codes to have a Toeplitz structure, which in effect means that the codes are time-invariant convolutional codes of growing constraint length, the authors demonstrate the existence of semi-infinite causal

linear codes for the binary-erasure and binary-symmetric channels. The erasure channel also affords an efficient decoding algorithm. This allows the selection of deterministic codes, in contrast to the random constructions considered in this paper. Efficient decoding of the family of codes considered herein was also considered in [25].

B. Outline

In Section II we review classic results on error exponents for fixed-block Slepian-Wolf source coding. In Section III we state the main result of the paper on error exponents for streaming Slepian-Wolf source coding and connect back to the form of the block coding exponents provided in Section II. In Section IV we present illustrative numerical results, including more detailed discussion of the example of Fig. 2. The theorems of Section III are proved in Sections V and VI. Section V begins by deriving results for point-to-point streaming source coding. This is the simplest case and provides insights into the nature of sequential source coding problem and associated error events. We show that the streaming error exponent is the same as the random block source coding error exponent. In Section V-E we consider point-to-point streaming source coding when side-information is available at the decoder. In Section VI we present the proof of the main result of the paper on the error exponents of distributed streaming source coding for correlated sources. For all three scenarios, point-to-point source coding, decoding with side-information, and distributed source coding, both maximum likelihood (ML) and universal decoding rules are studied. We defer the proofs of some lemmas to the appendices where, in addition we show that the error exponents achieved by the ML and universal decoders are, in fact, the same.

C. Notation

We use serifed-fonts, e.g., x to indicate sample values, and sans-serif, e.g., x, to indicate random variables. Bolded fonts are reserved to indicate sample or random vectors, e.g., $\mathbf{x} = x^n$ and $\mathbf{x} = \mathsf{x}^n$, respectively, where the vector length (n here) is understood from the context. Subsequences, e.g., $x_l, x_{l+1}, \ldots, x_n$ are denoted as x_l^n where $x_i^j \triangleq \emptyset$ if $i > j$. Distributions are indicated with lower-case p, e.g., x is distributed according to $p_{\mathsf{x}}(x)$. We use script font to denote sets, $\mathcal{X}$, $\mathcal{F}$, $\mathcal{W}$, etc., their cardinality by, e.g., $|\mathcal{X}|$, and reserve $\mathcal{E}$ and $\mathcal{D}$ to denote encoding and decoding functions, respectively. We use standard notation for types, see, e.g., [5]. Let $N(a; \mathbf{x})$ denote the number of symbols in the length-n vector $\mathbf{x}$ that take on value a. Then, $\mathbf{x}$ is of type P if $P(a) = N(a; \mathbf{x})/n$. The type-class, or set of length-n vectors of type P is denoted $\mathcal{T}_P$. A sequence $\mathbf{y}$ has conditional type V given $\mathbf{x}$ if $N(a, b; \mathbf{x}, \mathbf{y}) = N(a; \mathbf{x})V(b|a) = nP(a)V(b|a)$ for every a, b. The set of sequences $\mathbf{y}$ having conditional type V with respect to $\mathbf{x}$ is called the V-shell of $\mathbf{x}$ and is denoted by $\mathcal{T}_V(\mathbf{x})$. When considered together, the pair $(\mathbf{x}, \mathbf{y})$ is said to have joint type $V \times P$. We always use upper-case, e.g., P and V, to denote length-n types and conditional types. As we often discuss the types of subsequences we add a superscript notation to remind the reader of the length of the subsequence in question. If, for instance, the subsequence under consideration is x_l^n we write $x_l^n \in \mathcal{T}_{P^{n-l}}$. Similarly we use V^{n-l} for the conditional type of length-$(n - l + 1)$, and $V^{n-l} \times P^{n-l}$ for the joint type. Given a joint type $V \times P$, entropies and conditional entropies are denoted as $H(P)$ and $H(V|P)$, respectively. Alternately, the empirical joint entropy of a pair of sequences (x^n, y^n) is denoted $H(x^n, y^n)$. The entropy of a Bernoulli-p distribution is denoted as $H_B(p)$. Generally we assume the natural-base for our logarithms, expressing entropies in nats. The one exception is in Section IV where we use bits since one of our prominent examples is binary. The Kullback Leibler (KL) divergence between two distributions q and p is denoted by $D(q\|p)$. Finally, $|\cdot|^+$ is used as shorthand to denote $\max(\cdot, 0)$.

II. Background Results

In this section we review classical definitions and error exponent results for distributed block coding. In later sections we refer back to these results to contrast them with the results from the streaming framework.

In the classic block-coding Slepian-Wolf paradigm, length-N vectors $\mathbf{x}$ and $\mathbf{y}$ are observed by their respective encoders before communication commences. In this situation a rate-(R_x, R_y) length-N block source code consists of an encoder-decoder triplet $(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)$:

Definition 1: A randomized length-N rate-(R_x, R_y) block encoder-decoder triplet $(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)$ is a set of maps

$$\begin{aligned}
&\mathcal{E}_N^x : \mathcal{X}^N \to \{0,1\}^{NR_x}, \text{ e.g., } \mathcal{E}_N^x(x^N) = a^{NR_x}\\
&\mathcal{E}_N^y : \mathcal{Y}^N \to \{0,1\}^{NR_y}, \text{ e.g., } \mathcal{E}_N^y(y^N) = b^{NR_y}\\
&\mathcal{D}_N : \{0,1\}^{NR_x} \times \{0,1\}^{NR_y} \to \mathcal{X}^n \times \mathcal{Y}^n,\\
&\qquad\qquad \text{e.g., } \mathcal{D}_N(a^{NR_x}, b^{NR_y}) = (\hat{x}^N, \hat{y}^N)
\end{aligned}$$

where common randomness, shared between the encoders and the decoder is assumed. This allows us to randomize the mappings independently of the source sequences.

While we state Definition 1 only for Slepian-Wolf coding, it immediately specializes to source coding with decoder side information (dropping the $\mathcal{E}_N^y$ and revealing y^N to the decoder), and point-to-point source coding without side information (dropping the $\mathcal{E}_N^y$).

The standard error probability considered in Slepian-Wolf coding is the joint error probability, $\Pr[(x^N, y^N) \neq (\hat{x}^N, \hat{y}^N)] = \Pr[(x^N, y^N) \neq \mathcal{D}_N(\mathcal{E}_N^x(x^N), \mathcal{E}_N^y(y^N))]$. In this paper we also consider the marginal error events $\Pr[x^N \neq \hat{x}^N]$ and $\Pr[y^N \neq \hat{y}^N]$. Distinguishing between these events is of interest in applications where $\mathbf{x}$ and $\mathbf{y}$ are decoded jointly, but used individually. All probabilities are taken over the random source vectors as well as the randomized mappings. A joint error exponent E is said to be achievable if there exists a family of rate-(R_x, R_y) encoders and decoders $\{(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)\}$, indexed by N, such that

$$\lim_{N\to\infty} -\frac{1}{N} \log \Pr[(x^N, y^N) \neq (\hat{x}^N, \hat{y}^N)] \geq E. \tag{1}$$

Similarly, a marginal exponent E is achievable for source x^N if

$$\lim_{N\to\infty} -\frac{1}{N} \log \Pr[x^N \neq \hat{x}^N] \geq E. \tag{2}$$

In this paper, we study random source vectors $(\mathbf{x}, \mathbf{y})$ that are i.i.d. across time but may have dependencies at any given time:

$$p_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} p_{x,y}(x_i, y_i).$$

For such i.i.d. sources, upper and lower bounds on the achievable error exponents are derived in [5], [12], [16]. These results are summarized by the following theorems.

Theorem 1: Given rate pair (R_x, R_y), there exists a randomized encoder-decoder triplet (per Definition 1) that satisfy the following three decoding criteria:

(i) For all $E < E_{bl,x}(R_x, R_y)$, there is a constant $K > 0$ such that $\Pr[\hat{x}^N \neq x^N] \leq K \exp\{-NE\}$ where

$$E_{bl,x}(R_x, R_y) = \min\left\{ \sup_{0\leq\rho\leq 1} E_{x|y}(R_x, \rho), \sup_{0\leq\rho\leq 1} E_{xy}(R_x, R_y, \rho)\right\}. \quad (3)$$

(ii) For all $E < E_{bl,y}(R_x, R_y)$ there is a constant $K > 0$ such that $\Pr[\hat{y}^N \neq y^N] \leq K \exp\{-NE\}$ where

$$E_{bl,y}(R_x, R_y) = \min\left\{ \sup_{0\leq\rho\leq 1} E_{y|x}(R_y, \rho), \sup_{0\leq\rho\leq 1} E_{xy}(R_x, R_y, \rho)\right\}. \quad (4)$$

(iii) For all $E < E_{bl,xy}(R_x, R_y)$ there is a constant $K > 0$ such that $\Pr[(\hat{x}^N, \hat{y}^N) \neq (x^N, y^N)] \leq K \exp\{-NE\}$ where

$$E_{bl,xy}(R_x, R_y) = \min\left\{ \sup_{0\leq\rho\leq 1} E_{x|y}(R_x, \rho), \sup_{0\leq\rho\leq 1} E_{y|x}(R_y, \rho), \sup_{0\leq\rho\leq 1} E_{xy}(R_x, R_y, \rho)\right\}. \quad (5)$$

In the above,

$$E_{xy}(R_x, R_y, \rho) = \rho(R_x + R_y) - \log\Big[\sum_{x,y} p_{xy}(x, y)^{\frac{1}{1+\rho}}\Big]^{1+\rho} \quad (6)$$

$$E_{x|y}(R_x, \rho) = \rho R_x - \log\Big[\sum_{y}\Big[\sum_{x} p_{xy}(x, y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] \quad (7)$$

$$E_{y|x}(R_y, \rho) = \rho R_y - \log\Big[\sum_{x}\Big[\sum_{y} p_{xy}(x, y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big]. \quad (8)$$

As long as (R_x, R_y) is in the interior of the achievable Slepian-Wolf region, i.e., $R_x > H(x|y)$, $R_y > H(y|x)$ and $R_x + R_y > H(x, y)$, cf. [4], [27], all the above exponents are positive. Upper bounds on the error exponents are provided in [5], and match the lower bounds when the rate pair (R_x, R_y) is within, but close to the boundary of, the achievable region. This is analogous to the high-rate regime in channel coding where the random coding and sphere-packing bounds match.

Theorem 1 can be used to generate bounds on the exponent for source coding with decoder side information (i.e., $\mathbf{y}$ observed at the decoder), and for source coding without side information (i.e., $\mathbf{y}$ is a constant). These corollaries will serve as a basis for comparison as we build toward the complete solution for streaming Slepian-Wolf systems.

Corollary 1: Consider a Slepian-Wolf problem where $\mathbf{y}$ is known by the decoder. Given a rate R_x, then for all

$$E < \sup_{0\leq\rho\leq 1} \rho R_x - \log\Big[\sum_{y}\Big[\sum_{x} p_{xy}(x, y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] \quad (9)$$

there exists a family of randomized encoder-decoder mappings as defined in Definition 1 such that (2) is satisfied.

The proof of Corollary 1 follows from Theorem 1 by letting R_y be arbitrarily large. Note that the exponent in (9) is identical to $E_{x|y}(R_x, \rho)$ in (6), which given an operational meaning to that exponent. That exponent bounds the event that $\mathbf{x}$ is decoded incorrectly while $\mathbf{y}$ is decoded correctly.

Next let $\mathbf{y}$ be deterministic, e.g., $p_{x,y}(x, y) = p_{x|y}(x|y)1[y = a]$ for some $a \in \mathcal{Y}$ where $1[\cdot]$ is the indicator function. Then it follows that $H(x) = 0$, $H(x|y) = H(x)$ and, specializing the form of $E_{x|y}(R_x, \rho)$ to this distribution, we get the following random-coding bound for the point-to-point case of a single source $\mathbf{x}$.

Corollary 2: Consider a Slepian-Wolf problem where $\mathbf{y}$ is deterministic, i.e., $\mathbf{y} = \mathbf{y}$. Given a rate R_x, then for all

$$E < \sup_{0\leq\rho\leq 1} \rho R_x - \log\Big[\sum_{x} p_x(x)^{\frac{1}{1+\rho}}\Big]^{1+\rho} \quad (10)$$

there exists a family of randomized encoder-decoder triplet as defined in Definition 1 such that (2) is satisfied.

Gallager [12] and Koshelev [16] initiated the study of the error exponents of ML decoding for Slepian-Wolf systems, Gallager for source coding with decoder side information, and Koshelev for the two-encoder Slepian-Wolf problem. The joint decoding bound (5) is from [16] where (in the case of ML decoding considered therein) the constant $K = 1$. Koshelev did not consider the marginal exponents (3) and (4), but those can be extracted immediately from his derivation. As might be guessed from the discussion following Corollary 1, Koshelev partitions the joint error event (1) into three constituent events: (a) both $\hat{x}^N$ and $\hat{y}^N$ are erroneous, (b) only $\hat{x}^N$ is erroneous, (c) only $\hat{y}^N$ is erroneous. Respectively, the exponents bounding each of these events are given in (6)–(8). By ignoring either of the latter two events one get the marginal error bounds. For example, ignoring event (c) and accounting for events (a) and (b) leads to a bound on the event that only $\hat{x}^N$ is erroneous, and to the error exponent of (3).

It is well known in the literature [5] that the results of Theorem 1 and Corollaries 1 and 2 can be achieved by universal decoders as well as ML decoders. Universal decoding results are often derived using the methods of types, e.g., in [5]. The "Csiszár-style" exponents of [5] take a different form from the "Gallager-style" form of the exponents given in this section, due to the use of type-based arguments. The equivalence of the two forms of the exponents for these problems is a classic result. See, e.g., [5, pg. 44] exercise 13 and [5, pg. 192] exercise 23.

III. Main Results

In this section we present the main results of the paper. We define the functionality of streaming source coding for both point-to-point and distributed systems. We present results

for both maximum likelihood (ML) and universal decoding. The error exponents achieved are equal for both. We compare the forms of the streaming exponents with their block coding counterparts and in Section IV illustrate the differences through numerical examples. Proofs of the results are provided in Sections V and VI, while we defer to the appendices proofs not needed to understand the fundamental differences between block and streaming coding.

A. Code Definitions and Error Events for Streaming Systems

We start by defining sequential fixed-rate encoder/decoder pairs for streaming source coding systems. As we comment, in this paper we exclusively focus on encoders that employ random binning.

Definition 2: A randomized sequential rate-(R_x, R_y) encoder-decoder triplet $(\{\mathcal{E}_j^x\}, \{\mathcal{E}_j^y\}, \{\mathcal{D}_j\})$ is a sequence of mappings, $\{\mathcal{E}_j^x\}, j = 1, 2, \ldots$, $\{\mathcal{E}_j^y\}, j = 1, 2, \ldots$ and $\{\mathcal{D}_j\}, j = 1, 2, \ldots$ such that

$$\mathcal{E}_j^x : \mathcal{X}^j \longrightarrow \{0,1\}^{\lfloor jR_x \rfloor - \lfloor (j-1)R_x \rfloor}, \tag{11}$$

$$\mathcal{E}_j^y : \mathcal{Y}^j \longrightarrow \{0,1\}^{\lfloor jR_y \rfloor - \lfloor (j-1)R_x \rfloor}. \tag{12}$$

for example,

$$\mathcal{E}_j^x(x^j) = a_{\lfloor (j-1)R_x \rfloor + 1}^{\lfloor jR_x \rfloor},$$
$$\mathcal{E}_j^y(y^j) = b_{\lfloor (j-1)R_y \rfloor + 1}^{\lfloor jR_y \rfloor},$$

and where if $\lfloor (j-1)R_x \rfloor + 1 > \lfloor jR_x \rfloor$ the null sequence is produced. Common randomness, shared between encoders and decoder, is assumed. This allows us to randomize the mappings independently of the source sequence. Finally, the decoder mapping

$$\mathcal{D}_j : \{0,1\}^{\lfloor jR_x \rfloor} \times \{0,1\}^{\lfloor jR_y \rfloor} \longrightarrow \hat{\mathcal{X}}^j \times \hat{\mathcal{Y}}^j, \text{ e.g.,}$$
$$\mathcal{D}_j(a^{\lfloor jR_x \rfloor}, b^{\lfloor jR_y \rfloor}) = (\hat{x}^j(j), \hat{y}^j(j)).$$

At each time j the decoder $\mathcal{D}_j$ outputs estimates of all the source symbols that have entered the encoder by time j.

Note that sometimes we will allow an extra "failure" symbol "?" so that $\hat{\mathcal{X}} = \mathcal{X} \cup \{?\}$. In understanding these definitions it may help to recall the discussion of linear constructions in Section I-A and the lower-triangular nature of the parity-check matrix of those constructions.

As for block coding, while we state Definition 2 only for Slepian-Wolf coding, it immediately specializes to source coding with decoder side information (dropping the $\mathcal{E}_N^y$ and revealing y^N to the decoder), and point-to-point source coding without side information (dropping $\mathcal{E}_N^y$ and y^N completely).

In this paper, the sequential encoding maps will always work by assigning random "parity bits" in a causal manner to the observed source sequence. That is, the bits generated in (11)-(12), are i.i.d. Bernoulli-(0.5). Since parity bits are assigned causally, if two source sequences share the same length-l prefix, then their first $\lfloor lR_x \rfloor$ parity bits must match. Subsequent parities are drawn independently. Such a sequential coding strategy is the source-coding parallel to tree and convolutional codes used for channel coding [11]. In fact, we call these "parity bits" as they can be generated using an infinite constraint-length time-varying randomized convolutional code.

We will often restrict our attention to the set of source sequences that are compatible with the received parities up to time n. Given that $x^n = x^n$ this set is denoted as

$$\mathcal{B}_x(x^n) = \{\tilde{x}^n \in \mathcal{X}^n : \mathcal{E}_j^x(\tilde{x}^j) = \mathcal{E}_j^x(x^j), \quad j = 1, 2, \ldots, n\}. \tag{13}$$

An analogous definition holds for $\mathcal{B}_y(y^n)$.

We define the pair of source estimates at time n as $(\hat{x}^n, \hat{y}^n) = \mathcal{D}_n(\prod_{j=1}^n \mathcal{E}_j^x, \prod_{j=1}^n \mathcal{E}_j^y)$, where $\prod_{j=1}^n \mathcal{E}_j^x$ indicates the full nR_x bit stream from encoder x up to time n. We use $(\hat{x}^{n-\Delta}, \hat{y}^{n-\Delta})$ to indicate the first $n - \Delta$ symbols of each estimate, where for conciseness of notation both the estimate time, n, and the decoding delay, Δ, are indicated in the superscript. With these definitions the two marginal error probabilities are

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \text{ and } \Pr[\hat{y}^{n-\Delta} \neq y^{n-\Delta}].$$

A pair of exponents $E_x > 0$ and $E_y > 0$ is said to be achievable if there exists a family of rate-(R_x, R_y) encoders and decoders $\{(\mathcal{E}_j^x, \mathcal{E}_j^y, \mathcal{D}_j)\}$ such that

$$\lim_{\Delta \to \infty} \inf_{n > \Delta} -\frac{1}{\Delta} \log \Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \geq E_x, \tag{14}$$

$$\lim_{\Delta \to \infty} \inf_{n > \Delta} -\frac{1}{\Delta} \log \Pr[\hat{y}^{n-\Delta} \neq y^{n-\Delta}] \geq E_y. \tag{15}$$

In contrast to the block-coding error event of (1) this error exponent is in delay, Δ, rather than total observation time, n. While the definitions of the exponents (14)–(15) and of (1) are asymptotic in nature, the error bounds stated in the theorems hold for finite n and Δ. Finally, we note that, as in the block coding case, the error exponent of the joint error event can be found by taking the minimum of the individual exponents, i.e.,

$$\lim_{\Delta \to \infty} \inf_{n > \Delta} -\frac{1}{\Delta} \log \Pr[(\hat{x}^{n-\Delta}, \hat{y}^{n-\Delta}) \neq (x^{n-\Delta}, y^{n-\Delta})] \geq \min\{E_x, E_y\}.$$

We can now see why the use of randomized maps is important. This is due to the infinite operating horizon of our system and the fact that we require exponential decay in error probability *at all times* and *for all delays*. We will show that the desired performance can be achieved over the ensemble of tree codes through the use of commonly randomized encoders and decoders. However, the standard argument that because the ensemble of codes satisfy some measure of performance therefore a single code must exist that does also cannot be applied as there is now a countable number of measures of performance (all times and all delays).

B. Point-to-Point Streaming

Our first results concern streaming coding in the point-to-point setting. The first theorem provides achievable bounds on the random coding error exponents both for ML and universal decoding.

Theorem 2: Given any rate R, there exist both ML and universal randomized sequential rate-R point-to-point encoder-decoder pairs (per the specialization of Definition 2) such that

for all $E < E_{pt,x}(R)$ there is a constant $K > 0$ such that $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{pt,x}(R) = \sup_{0\leq\rho\leq 1} \rho R - \log\Big[\sum_x p_X(x)^{\frac{1}{1+\rho}}\Big]^{1+\rho} \quad (16)$$

$$= \inf_{\bar{x}} D(p_{\bar{x}} \| p_X) + |R - H(\bar{x})|^+. \quad (17)$$

where in (17) $\bar{x}$ is a random variable on $\mathcal{X}$ with distribution $p_{\bar{x}}$ and entropy $H(\bar{x})$.

Proof: In Section V-C a Gallager-style analysis of ML decoding yields the form of the exponent specified in (16). This analysis is the source-coding parallel to the traditional one for convolutional channel codes. In Section V-D a types-based analysis of a novel universal decoder yields the form of the exponent specified in (17). Here, the crucial issue that must be side stepped is the non-additivity of empirical entropy. The equality of the two forms of the exponent in (16) and (17) is a classic result. For example, see [5, pg. 44] exercise 13. ■

The error exponent of Theorem 2 equals the random source coding exponent for block-coding (10). The main difference in the formulation is that the error probability in a streaming system decays with delay Δ rather than block length N. For any fixed source symbol with time index j, as time progresses ($n \to \infty$) the delay $\Delta = n - j$ also increases without bound. Thus all symbols are eventually recovered with probability one.

A companion upper-bound to Theorem 2 is proved in [2]. We state the theorem next, sketch the proof, and refer the reader to [2] for details.

Theorem 3: Any (randomized or deterministic) sequential rate-R point-to-point encoder-decoder pair satisfies

$$\lim_{\Delta\to\infty} \sup_{n>\Delta} -\frac{1}{\Delta} \log \Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq \inf_{\alpha>0} \frac{1}{\alpha} E^{up}_{pt,x}((1+\alpha)R). \quad (18)$$

where

$$E^{up}_{pt,x}(R) = \sup_{\rho\geq 0} \rho R - \log\Big[\sum_x p_X(x)^{\frac{1}{1+\rho}}\Big]^{1+\rho}$$

$$= \inf_{\bar{x}: H(\bar{x})>R} D(p_{\bar{x}} \| p_X). \quad (19)$$

Proof sketch: Consider the decoding of the tth symbol x_t at time $t + \Delta$ for any arbitrary t by a delay-constrained decoder. Consider an encoder/decoder pair aided in the following ways: (i) we tell the decoder symbols $x_{t+1}, \ldots, x_{t+\Delta}$; (ii) we tell the decoder symbols $x_0, \ldots x_{t-L}$ for some L to be determined; (iii) we tell the encoder the L symbols $x_{t-L+1}, \ldots, x_t$ all at time $t - L + 1$; (iv) we don't require the decoder to decode any of the symbols $x_{t-L+1}, \ldots, x_t$ until time $t + \Delta$. A pair so enabled is *strictly* more powerful than our usual delay-constrained encoder/decoder pair since these stronger pairs can emulate our usual pairs. Further, since

$$\max_{i\in\{t-L+1,\ldots,t\}} \Pr[\hat{x}_i \neq x_i] \geq \frac{1}{L} \Pr[\cup_{i=t-L+1}^{t} (\hat{x}_i \neq x_i)]$$

and t is arbitrary, by lower bounding the block-error-probability of the more powerful encoder/decoder pair, we gain a lower bound on the error probability of our usual pairs.

Note that by (i) and (ii) we are able to ignore the distance past (before time $t-L+1$, where L is yet to be determined) and the future. Further by (iii) and (iv) we have transformed the problem into an equivalent block coding problem. The length of this block code is L. The rate is $(L+\Delta)R/L = (1+\Delta/L)R$. The block error probability of such a block code is lower bounded by classic results for block coding, cf. [5]. Since the classic bounds hold both for deterministic and random coding, the current bound also holds for both deterministic and random strategies. Namely,

$$\Pr[\hat{x}_{t-L+1}^{t} \neq x_{t-L+1}^{t}] \dot{\geq} \exp\Big\{-LE^{up}_{pt,x}\Big(\Big(1+\frac{\Delta}{L}\Big)R\Big)\Big\},$$

where $E^{up}_{pt,x}(\cdot)$ is the classic source coding bound for block codes specified in (19). By $\dot{\geq}$ we mean the relation holds to the first order in the exponent [4] (if a good source code is used $\dot{\geq}$ can be replaced by $\doteq$). We state the result in this way to suppress non-exponential factors that will do affect the exponent, thereby simplifying the presentation. Recalling that L is a free parameter yet to be specified, we get the tightest bound by finding the (worst-case) L that maximizes the bound:

$$\max_L \exp\Big\{-LE^{up}_{pt,x}\Big(\Big(1+\frac{\Delta}{L}\Big)R\Big)\Big\}$$

$$= \exp\Big\{-\Delta \inf_{\alpha>0} \frac{1}{\alpha} E^{up}_{pt,x}\Big(\Big(1+\alpha\Big)R\Big)\Big\}$$

Taking the log, normalizing, and recalling that this bound holds for all Δ we get the result in (18). ■

The idea of the proof is that there is some atypicality event that starts L samples in the past. The worst-case time in the past is $L^* = \Delta/\alpha^*$ where α^* is the optimizer. Even if we ignore the distant past and the future symbols and concentrate all our resources, over the entire interval of length $L+\Delta$ from time $t-L+1$ to time $t+\Delta$, on correcting this error, correction is not possible. By converting the problem into an equivalent block coding problem we are able to leverage existing results for such systems.

We note the the derivation assumes that $t > \Delta$ and $t > L$. However, the first isn't restrictive since we are only interested in $\Delta \leq t$. The second isn't restrictive since if the optimizing $L^* > t$ we can add dummy symbols spanning time $t - L^*$ to time 0. Forcing the decoder to decode these will simply worsen the performance, further lowering the lower bound.

C. Streaming With Decoder Side Information

Our result for distributed streaming source coding when the side information is observed at the decoder, but not the encoder, is encapsulated in the following theorem:

Theorem 4: Given any rate R, there exist both ML and universal randomized sequential rate-R source coding with decoder side-information encoder-decoder pairs (per the specialization of Definition 2) such that for all $E < E_{si}(R)$ there is a constant $K > 0$ such that $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{si}(R) = \sup_{0\leq\rho\leq 1} \rho R - \log\Big[\sum_y \Big[\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] \quad (20)$$

$$= \inf_{\bar{x},\bar{y}} D(p_{\bar{x},\bar{y}} \| p_{x,y}) + |R - H(\bar{x}|\bar{y})|^+, \quad (21)$$

and $(\bar{x}, \bar{y})$ are random variables with joint distribution $p_{\bar{x},\bar{y}}$ and $H(\bar{x}|\bar{y})$ is their conditional entropy.

Similar to the point-to-point case in Theorem 2, the error exponent of Theorem 4 equals its random block-coding counterpart (9). Similarly, (20) and (21) can be shown to be equal. We do not prove this equivalence herein but, as a first step, the interested reader could consider [5, pg. 192] exercise 23. We sketch the proof of this theorem in Section V-E, which requires only small modifications of the techniques used to prove Theorem 2.

The following companion upper bound is proved in [2]:

Theorem 5: Any (randomized or deterministic) sequential rate-R source coding with decoder side-information encoder-decoder pair satisfies

$$\lim_{\Delta\to\infty} \sup_{n>\Delta} -\frac{1}{\Delta} \log \Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq E_{si}^{up}(R)$$

where

$$E_{si}^{up}(R) = \min \left\{ \begin{array}{l} \inf\limits_{\substack{\bar{x},\,\bar{y},\,\alpha \geq 1 \text{ s.t.} \\ H(\bar{x}|\bar{y}) \geq (1+\alpha)R}} \frac{1}{\alpha} D(p_{\bar{x},\bar{y}} \| p_{x,y}), \\ \inf\limits_{\substack{\bar{x},\,\bar{y},\,0 \leq \alpha \leq 1 \text{ s.t.} \\ H(\bar{x}|\bar{y}) \geq (1+\alpha)R}} \frac{1-\alpha}{\alpha} D(p_{\bar{x}} \| p_x) + D(p_{\bar{x},\bar{y}} \| p_{x,y}) \end{array} \right\}, \quad (22)$$

and $(\bar{x}, \bar{y})$ are random variables with joint distribution $p_{\bar{x},\bar{y}}$ and $H(\bar{x}|\bar{y})$ is their conditional entropy.

The proof approach here is similar to that sketched for Theorem 3. The main difference is that there are now two possible sources of error: there is the possibility of atypicality in the x-source and there is possibility of joint atypicality in the (x, y)-source pair. Either one type can dominate (as in the first case of (22) where joint atypicality dominates), or a combination of errors can occur (as in the second case of (22)). When the encoder can tell that such atypicality is taking place the encoder can take remedial action, e.g., by momentarily ignoring the recently arrived symbols (whose deadline is not yet close) and focusing resources (parity bits) on the symbols that are behaving atypically. Note that the rate is fixed throughout, it is simply a question of when each source symbol observed is allowed to start to affect the encoded parity symbols. In certain situations it is not possible for the encoder to detect that such atypicality is taking place. An example of this is a uniformly distributed source and a conditional relationship $p_{y|x}(y|x)$ that corresponds to a symmetric channel. This is akin to the condition given in in Remark 10 of [15] (mentioned in Section I-A) that characterizes when variable-rate source coding with decoder side information is no better than fixed-rate source coding, i.e., that

$$-\log p_x(x) - D(p_{y|x}(\cdot|x) \| p_y) \quad (23)$$

is constant in $x \in \mathcal{X}$. We also remark that when side information is also available at the encoder the upper bound on the exponent is, naturally, increased. In the latter setting one can apply the exponent of (19) to each conditional probability with an average across the various side information symbols. See [2] for details.

As an example of a special case where the upper bound reduces to a simpler form (equivalent to an upper bound for block coding) is when the side information y is uniformly distributed and $x = y \oplus e$ where e is independent of y (and $|\mathcal{X}| = |\mathcal{Y}| = |\mathcal{E}|$ so we can define the addition operator). Note that for this case (23) is constant. For such situations (22) simplifies to

$$E_{si}^{up}(R) = \inf_{\bar{x},\bar{y}: H(\bar{x}|\bar{y}) \geq R} D(p_{\bar{x},\bar{y}} \| p_{x,y}). \quad (24)$$

For such sources, e.g., a doubly-symmetric binary source, the upper bound of (24) and lower bounds of (20) and (21) match at rates close to the conditional entropy $H(x|y)$. We direct the reader to Appendix I of [2] where this example is fully developed.

D. Distributed Coding of Streaming Sources

In contrast to streaming point-to-point coding and streaming source coding with decoder side information, our results for the general case of streaming Slepian-Wolf coding with two separate encoders results in achievable error exponents that differ from their block coding counterparts.

Theorem 6: Given any rate pair (R_x, R_y), there exist both ML and universal randomized sequential rate-(R_x, R_y) Slepian-Wolf encoder-decoder triplets (per Definition 2) that satisfy the following three criteria:

(i) For all $E < E_{st,x}(R_x, R_y)$, there is a constant $K > 0$ such that $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{st,x}(R_x, R_y) = \min \left\{ \begin{array}{l} \inf\limits_{\gamma \in [0,1]} E_x(R_x, R_y, \gamma), \\ \inf\limits_{\gamma \in [0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma) \end{array} \right\}. \quad (25)$$

(ii) For all $E < E_{st,y}(R_x, R_y)$ there is a constant $K > 0$ such that $\Pr[\hat{y}^{n-\Delta} \neq y^{n-\Delta}] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{st,y}(R_x, R_y) = \min \left\{ \begin{array}{l} \inf\limits_{\gamma \in [0,1]} \frac{1}{1-\gamma} E_x(R_x, R_y, \gamma), \\ \inf\limits_{\gamma \in [0,1]} E_y(R_x, R_y, \gamma) \end{array} \right\}. \quad (26)$$

(iii) For all $E < E_{st,xy}(R_x, R_y)$ there is a constant $K > 0$ such that $\Pr[(\hat{x}^{n-\Delta}, \hat{y}^{n-\Delta}) \neq (x^{n-\Delta}, y^{n-\Delta})] \leq K \exp\{-\Delta E\}$ for all $n, \Delta \geq 0$ where

$$E_{st,xy}(R_x, R_y) = \min \left\{ \begin{array}{l} \inf\limits_{\gamma \in [0,1]} E_x(R_x, R_y, \gamma), \quad (27) \\ \inf\limits_{\gamma \in [0,1]} E_y(R_x, R_y, \gamma) \end{array} \right\}. \quad (28)$$

There are two alternate, but equivalent, ways to specify the above error exponents. The first is the "Gallager-style"

$$\begin{aligned} &E_x(R_x, R_y, \gamma) \\ &\quad = \sup_{\rho\in[0,1]} [\gamma E_{x|y}(R_x, \rho) + (1-\gamma)E_{xy}(R_x, R_y, \rho)] \\ &E_y(R_x, R_y, \gamma) \\ &\quad = \sup_{\rho\in[0,1]} [\gamma E_{y|x}(R_y, \rho) + (1-\gamma)E_{xy}(R_x, R_y, \rho)], \end{aligned} \quad (29)$$

where $E_{xy}(\cdot,\cdot,\cdot)$, $E_{x|y}(\cdot,\cdot)$, and $E_{y|x}(\cdot,\cdot)$ are defined as in (6)–(8), repeated here for convenience:

$$E_{xy}(R_x, R_y, \rho) = \rho(R_x + R_y) - \log\Big[\sum_{x,y} p_{xy}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}$$

$$E_{x|y}(R_x, \rho) = \rho R_x - \log\Big[\sum_{y}\Big[\sum_{x} p_{xy}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big]$$

$$E_{y|x}(R_y, \rho) = \rho R_y - \log\Big[\sum_{x}\Big[\sum_{y} p_{xy}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big].$$

Alternately, the second "Csiszár-style" form of the exponents is

$$\begin{aligned} &E_x(R_x, R_y, \gamma) \\ &\quad = \inf_{\tilde{x},\tilde{y},\bar{x},\bar{y}} \gamma D(p_{\tilde{x},\tilde{y}}\|p_{x,y}) + (1-\gamma)D(p_{\bar{x},\bar{y}}\|p_{x,y}) \\ &\qquad + \Big|\gamma[R_x - H(\tilde{x}|\tilde{y})] + (1-\gamma)[R_x + R_y - H(\bar{x},\bar{y})]\Big|^+ \\ &E_y(R_x, R_y, \gamma) \\ &\quad = \inf_{\tilde{x},\tilde{y},\bar{x},\bar{y}} \gamma D(p_{\tilde{x},\tilde{y}}\|p_{x,y}) + (1-\gamma)D(p_{\bar{x},\bar{y}}\|p_{x,y}) \\ &\quad + \Big|\gamma[R_y - H(\tilde{y}|\tilde{x})] + (1-\gamma)[R_x + R_y - H(\bar{x},\bar{y})]\Big|^+, \end{aligned} \quad (30)$$

where the random variables $(\tilde{x},\tilde{y})$ and $(\bar{x},\bar{y})$ have joint distributions $p_{\tilde{x},\tilde{y}}$ and $p_{\bar{x},\bar{y}}$, respectively.

Proof: In Section VI-C a Gallager-style analysis of ML decoding yields the form of the exponents specified in (29). In Section VI-D a types-based analysis of a novel universal decoder yields the form of the exponent specified in (30). The equality of the two forms of the exponent is considered in Lemma 5, stated and proved in Appendix C. ■

We first note that the exponents are strictly positive for any rate-pair (R_x, R_y) within the Slepian-Wolf achievability region. This is easiest to see by considering the Csiszár-style results of (30). By separately considering the case $\gamma = 0$, $\gamma = 1$, and $0 < \gamma < 1$ one can confirm that there is always at least one term in $E_x(R_x, R_y, \gamma)$ and in $E_y(R_x, R_y, \gamma)$ that must be strictly positive.

It is revealing to compare the form of the exponents of block coding (3)–(5) with those of streaming (25)–(28). The streaming exponent contains an extra degree of freedom in the parameter γ. If γ were restricted to be either zero or one, then the block and streaming exponents would be the same. The minimization over γ where $0 \le \gamma \le 1$ results from a fundamental difference in the types of events that can cause errors in streaming as opposed to the block setting. In block coding there are only 3 error events (error in x^n, error in y^n, and errors in both), regardless of block length. In contrast, there are n^2 mutually exclusive error events when decoding x^n and y^n. These arise from the n^2 pairs of time indexes at which the error patterns can commence, i.e., $l, k \in \{1, 2, \ldots, n\}$ such that $\hat{x}^{l-1} = x^{l-1}$ and $\hat{y}^{k-1} = y^{k-1}$, but $\hat{x}_l \neq x_l$ and $\hat{y}_k \neq y_k$. We examine these error events in Section VI.

While the error exponents of block coding are always at least as large as the streaming exponents due to the lack of the γ parameter, direct comparison of the two is not really appropriate for two reasons. The first is that buffering delay is not accounted for in block coding. Streaming data must first be packetized into "chunks" of data of the appropriate length to which the block-encoding can be applied. Such packetization delay is not accounted for in the block coding exponents and, at worst, would double the delay on a particular symbols (those at the beginning of each block). The second reason is that in block coding the block length is fixed and therefore so is the resulting error probability. In the streaming context the error probability on the estimate of any fixed source symbol continues to decrease as time increments and the decoding delay Δ (for that particular symbol) increases.

Finally, as in the point-to-point setting, the two forms of the exponents in (29) and (30) are equal. But, due to new classes of error events possible in streaming, this equivalence now requires proof. This proof is provided in Lemma 5.

IV. Numerical Results

In this section we detail two examples. The first example is presented in part in Fig. 2 in the Introduction and helps us understand how source "burstiness" relates to the achievable error exponent in delay. For simplicity we present these results for lossless point-to-point streaming, i.e., Theorem 2. The second example illustrates the difference between the block-coding and streaming exponents for a simple distributed asymmetric binary source. In this section we express entropy in bits.

The source considered in the first example is a discrete memoryless source with alphabet $\{0, 1, \ldots, L\}$ where $p_X(0) = (1-\beta)$ and $p_X(x) = \beta/L$ for all $x \neq 0$. The β parameter specifies the "burstiness" of the source and the entropy of this source is $H(X) = H_B(\beta) + \beta \log L$. In Fig. 2 we consider $L = 4$ and $\beta = 0.5$, hence $H(X) = 2$ bits. Fig. 2 plots the trade-off between rate, delay, and probability of error for ML decoding. (We note that the results for universal decoding would differ little as they are equal to ML in an exponential sense.) As would be expected, the probability of decoding error drops both as a function of communication rate and delay.

The source of Fig. 2 is only mildly bursty, half the time it emits a 0 and half the time some other letter. In Fig. 3 we plot the error exponent of the same family of sources for a range of burst probabilities β and alphabet sizes $L+1$ where we hold the entropy constant at $H(X) = 2$ bits. As the source becomes more bursty (smaller β) we increase the alphabet size to maintain the equality $H(X) = 2 = H_B(\beta) + \beta \log L$. The figure shows that that the more bursty the source (smaller β and large L) the smaller the error exponent for any given rate.

Our second example illustrates a distributed source coding situation where the streaming and block coding error exponents differ. The reason for the difference is the new type of

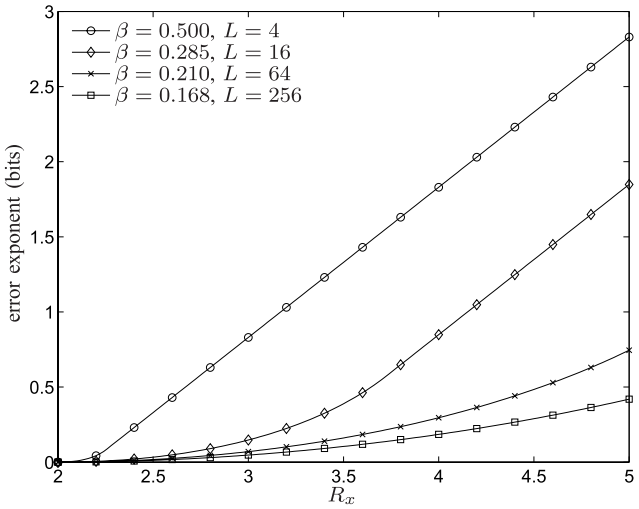

Fig. 3. The effect of burstiness on the error exponent as a function of the excess rate beyond the entropy. The source is 0 with probability $1-\beta$ and, with probability β, the source is uniformly distributed on $\{1, 2, \ldots, L\}$. We scale L with the burst probability β to hold the source entropy constant at 2 bits. A lower burst probability β means more variability in the instantaneous rate, the effect of which is a lowered exponent.

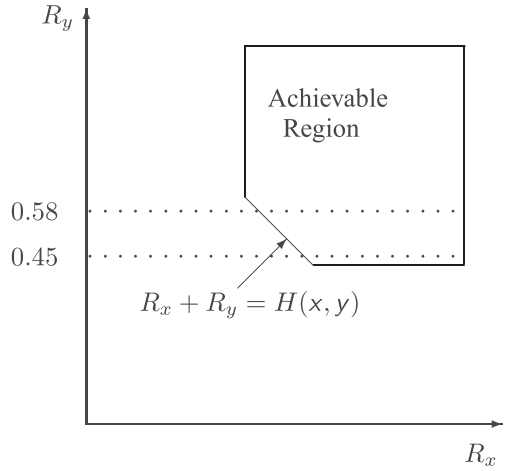

Fig. 4. Rate region for the asymmetric example source.

error event (reflected in the minimization over γ in Theorem 6) that can dominate in the distributed streaming setting. However, it turns out that when the distributed source has uniform symmetric marginals there is no gap between the streaming and block coding error exponents. Thus, we consider the following asymmetric example (asymmetric marginals and asymmetric channel relating x to y). The pair of sources x_i and y_i are binary i.i.d. sources where $p_{x,y}(0,0) = 0.1$, $p_{x,y}(0,1) = p_{x,y}(1,0) = 0.05$ and $p_{x,y}(1,1) = 0.8$. For this source $H(x) = H(y) = 0.61$ bits, $H(x|y) = H(y|x) = 0.42$ bits and $H(x,y) = 1.02$ bits. The Slepian-Wolf achievable rate region is shown in Fig. 4. We consider various error exponents for this source as a function of R_x where we keep R_y fixed. We consider both a low R_y-rate situation where $R_y = 0.45 = H(y|x) + 0.03$ bits and a high R_y-rate situation where $R_y = 0.58 = H(y) - 0.03$ bits.

Fig. 5 plots the streaming exponent $E_{st,x}(R_x, R_y)$ for source x from Theorem 6, the block coding exponent $E_{bl,x}(R_x, R_y)$ from Theorem 1, and the point-to-point exponent $E_{pt,x}(R_x)$ from Theorem 2. All are plotted as a function of R_x, and the first two for both $R_y = 0.45$ and $R_y = 0.58$. A note on the plots: since $E_{st,x}(R_x, R_y) = E_{bl,x}(R_x, R_y)$ for many choices of R_x and R_y, we choose to plot the block coding exponents with solid or dashed lines and the streaming exponents with circles or diamonds. Both are, of course, continuous functions of R_x. Our choice of plotting the streaming exponents at a discrete set of points was made purely to aid in making visual comparison between the exponents.

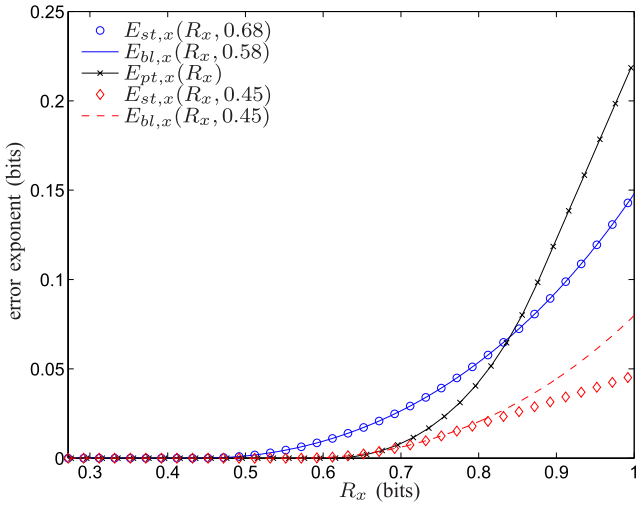

Fig. 5. Error exponents for x-source: streaming $E_{st,x}(R_x, R_y)$, block-coding $E_{bl,x}(R_x, R_y)$, and point-to-point source coding $E_{pt,x}(R_x)$ at two rates: $R_y = 0.68$ and $R_y = 0.45$ bits per sample.

There are a few observations to make about Fig. 5. Perhaps the most significant is that, in order to recover the x-source with the greatest likelihood, it can be better *not* to use joint decoding if R_y is too low. For example, when $R_y = 0.45$ and $R_x > 0.65$ bits, $E_{pt,x}(R_x)$ is larger than either $E_{st,x}(R_x, 0.45)$ or $E_{bl,x}(R_x, 0.45)$. This occurs because joint decoding errors are more likely due to atypical behavior of source y. Thus, it can be better to ignore the y-source and decode the x-source individually. As R_y is increased, e.g., to $R_y = 0.58$ bits, the information about the y-source is more reliable and the joint decoding exponents dominate that of point-to-point source coding without side information up to higher rates, about $R_x = 0.84$. The next observation to make is that the difference between the block and streaming error exponents is small and often zero. In Fig. 5 the difference between the two is only apparent about $R_x \simeq 0.75$ for the $R_y = 0.45$ case. To see more clearly where the streaming and block coding exponents differ, in Fig. 6 we plot the ratio $E_{bl,x}(R_x, R_y)/E_{st,x}(R_x, R_y)$. In this figure we see that at the higher rate $R_y = 0.58$ the exponents are the same for the entire range.

Figs. 7 and 8 plot the corresponding results for $E_{st,y}(R_x, R_y)$, $E_{bl,y}(R_x, R_y)$, and $E_{pt,y}(R_y)$ for $R_y = 0.45$ and $R_y = 0.58$. Note that $E_{pt,y}(0.45) = E_{pt,y}(0.58) = 0$ since both rates are below $H(y) = 0.61$ bits and $E_{pt,y}(0.71)$ is constant at about 0.01. Thus, joint decoding is required to get any positive exponent on the y-source. Next note that when R_x is sufficiently high, the the error exponent for y saturates

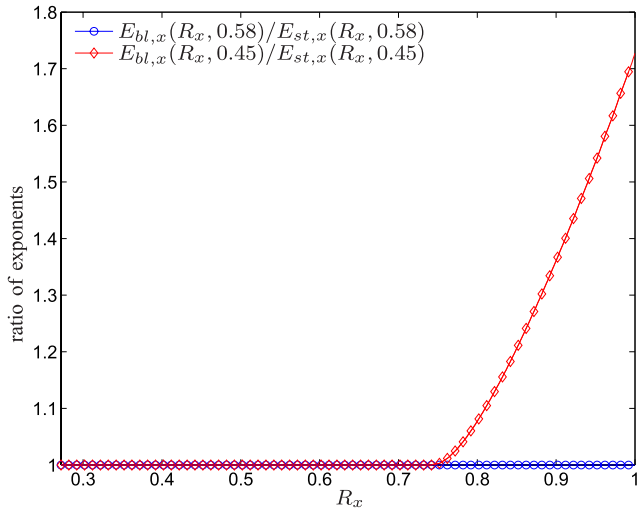

Fig. 6. Ratio of block-coding to streaming exponents for source-x. The block coding exponent is always at least as large due to the extra possibility error events in the streaming setting.

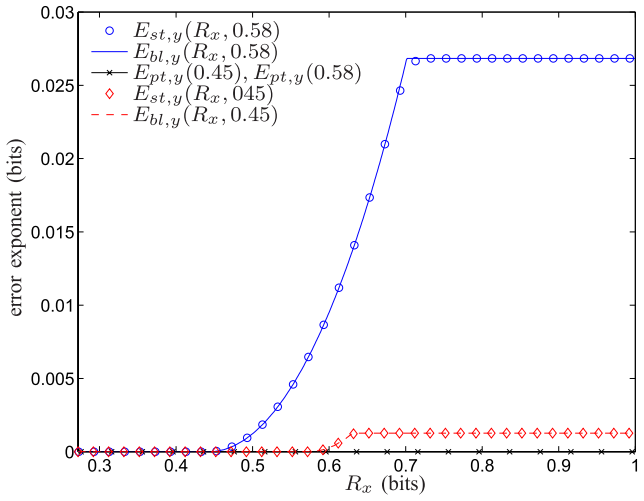

Fig. 7. Error exponents for y-source: streaming $E_{st,y}(R_x, R_y)$, block-coding $E_{bl,y}(R_x, R_y)$, and point-to-point source coding $E_{pt,y}(R_y)$ at two rates: $R_y = 0.45$ and $R_y = 0.58$ bits per sample. Note that $E_{pt,y}(0.45) = E_{pt,y}(0.58) = 0$.

to what it would be if source x were known perfectly to the decoder. Recalling the discussion in Sections II and III, this is the contribution of the $E_{y|x}(R_y, \rho)$ term to the exponents in (4) and (26). As before, to help visualize the difference between the block and streaming exponents, in Fig. 8 we plot the ratio $E_{bl,y}(R_x, R_y)/E_{st,y}(R_x, R_y)$. In this plot we note a feature that didn't appear in Fig. 6; namely, that for certain ranges of R_x (that don't overlap) the ratio of the exponents is greater than one both for the high- and low-R_y examples.

V. Streaming Point-to-Point Coding via Sequential Random Binning

In this section we prove Theorems 2 and 4. While the emphasis of the paper is on distributed source coding, the strategy of causal random binning, the appropriate ML and universal decoders, and the associated analysis techniques, are most easily developed in the point-to-point context.

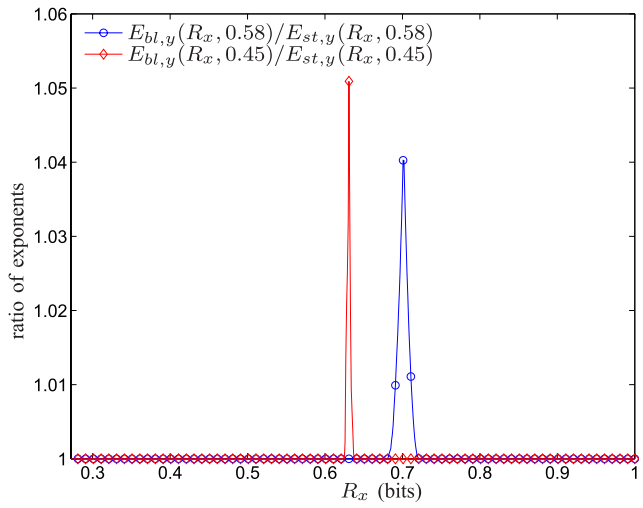

Fig. 8. Ratio of block-coding to streaming exponents for source-y.

A. Sequential Scoring Decoders

In this section we introduce the class of decoders used for streaming source coding and streaming source coding with decoder side information. Both ML and universal decoders can be cast as a type of decision-directed sequential scoring decoder, where different scoring functions are used in each case.

Definition 3: A sequential scoring decoder constructs its estimate in a sequential manner starting from $l = 1$ where

$$\hat{x}_l = \begin{cases} \bar{\mathbf{x}}_l & \text{if for some } \bar{\mathbf{x}} \in \mathcal{B}_x(\mathbf{x}) \text{ s.t } \bar{x}^{l-1} = \hat{x}^{l-1} \\ & S_l(\bar{\mathbf{x}}) \geq S_l(\tilde{\mathbf{x}}) \text{ for all } \tilde{\mathbf{x}} \in \mathcal{B}_x(\mathbf{x}) \\ & \text{s.t. } \tilde{x}^{l-1} = \hat{x}^{l-1}, \tilde{x}_l \neq \bar{x}_l \\ ? & \text{otherwise} \end{cases} \tag{31}$$

where we recall that $\mathcal{B}(\mathbf{x})$ is defined in (13) and where $S_l(\cdot)$ is a (possibly time-dependent) scoring function and the (failure) symbol "?" is included in case such a $\bar{\mathbf{x}}$ does not exists for some l. Randomly resolve any ties that occur.

Since the sequential scoring decoder is a decision-directed decoder, it considers as candidates only those sequences whose parities match the received bit stream up to time n, i.e., if the length-n source sequence is $\mathbf{x} = \mathbf{x}$ then the set of such candidates is $\{\bar{\mathbf{x}} \text{ s.t. } \bar{\mathbf{x}} \in \mathcal{B}_x(\mathbf{x})\}$. The lth symbol of the estimate, $\hat{x}_l$, is made with the estimates of the first $l-1$ symbols already fixed. One should note that as soon as the next set of parities arrive at the receiver, all symbols are estimated anew since n is now replaced by $n+1$ and $\mathcal{B}_x(x^{n+1})$ will be different from $\mathcal{B}_x(x^n)$.

For ML decoding case we use the scoring function

$$S_l(\bar{x}) = p_{X_l^n}(\bar{x}_l^n). \tag{32}$$

Note that this scoring function simply leads to the ML estimate $\hat{\mathbf{x}}_{ML} = \arg\max_{\bar{\mathbf{x}} \in \mathcal{B}_x(\mathbf{x})} p_{\mathbf{x}}(\bar{\mathbf{x}})$ being constructed in a sequential manner. This is the case since the decision regarding which of a pair of sequences is more likely depends only on which

sequence has the more likely suffix. Another way of saying this is that, if we were to consider the log-probability score $S_l = \log p_{X_l^n}(\bar{x}_l^n)$, then the score would be additive for i.i.d. sequences. Thus, we could equally have chosen $S_l(\bar{\mathbf{x}}) = p_{\mathbf{x}}(\bar{\mathbf{x}})$. On the other hand, since empirical entropy is not additive (think of a sequence of all 0s followed by a sequence of all 1s) the use of sequential scoring decoders will be more crucial in universal decoding.

For universal decoding we use the reciprocal of the empirical suffix-entropy as the score

$$S_l(\bar{x}) = 1/H(\bar{x}_l^n). \tag{33}$$

and term the resulting decoding the "minimum empirical suffix entropy decoder". The reason for using this decoder instead of the standard minimum empirical block-entropy decoder is because (due to the summing over type classes) the probability of error bound for the block-entropy decoder has a pre-multiplier term that grows polynomially in n. Since our bound on error probability will decay exponential in Δ, for n large, the polynomial can dominate. This would prevent us from deriving a bound on the probability of error that depends only upon the decoding delay Δ. Using the minimum empirical suffix-entropy decoder results in a term that grows polynomially only in Δ.

B. Error Analysis of Sequential Scoring Decoders

To show Theorem 2 we first develop the common core of the proof that applies to both ML and universal decoding. The proof strategy is as follows. A decoding error can only occur if there is some spurious source sequence $\tilde{x}^n$ that satisfies three conditions: (i) $\tilde{x}^n \in \mathcal{B}_x(x^n)$, i.e., it must be in the same bin (share the same parities) as x^n, (ii) $\tilde{x}_l \neq x_l$ for some $l \leq n-\Delta$, and (iii) for the time index l of event (ii) it must have a score at least as large as the correct sequence, i.e., $S_l(\tilde{x}^n) \geq S_l(x^n)$.

To help track condition (ii) and to keep notation compact we introduce a partition of all length-n source sequences $\tilde{x}^n \in \mathcal{X}^n$ into non-overlapping sets $\mathcal{F}_n(l, x^n)$ defined by the time index l of the first sample in which each sequence differs from the realized sequence x^n. Formally,

$$\mathcal{F}_n(l, x^n) = \{\tilde{x}^n \in \mathcal{X}^n | \tilde{x}^{l-1} = x^{l-1}, \tilde{x}_l \neq x_l\}, \tag{34}$$

where we define $\mathcal{F}_n(n+1, x^n) = \{x^n\}$, thus $\cup_{l=1}^{n+1} \mathcal{F}_n(l, x^n) = \mathcal{X}^n$.

With these definitions we rewrite the error probability as

$$\begin{aligned}
&\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \\
&= \sum_{x^n} \Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta} | x^n = x^n] p_{\mathbf{x}}(x^n) && (35)\\
&= \sum_{x^n} \sum_{l=1}^{n-\Delta} \Pr\big[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \\
&\qquad\qquad \text{s.t. } S_l(\tilde{x}^n) \geq S_l(x^n)\big] p_{\mathbf{x}}(x^n) && (36)\\
&= \sum_{l=1}^{n-\Delta} \Big\{ \sum_{x^n} \Pr\big[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \\
&\qquad\qquad \text{s.t. } S_l(\tilde{x}^n) \geq S_l(x^n)\big] p_{\mathbf{x}}(x^n) \Big\} && (37)
\end{aligned}$$

After conditioning on the realized source sequence in (35), the remaining randomness is only in the binning. In (36) the error event is decomposed into mutually exclusive events based on the discussion of conditions (i)-(iii) above, and the partitioning of all length-n sources sequences into the sets $\mathcal{F}_n(l, x^n)$. Finally, defining

$$p_n(l) = \sum_{x^n} \Pr\big[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } S_l(\tilde{x}^n) \geq S_l(x^n)\big] p_{\mathbf{x}}(x^n). \tag{38}$$

and substituting the results into (37) yields the relation

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] = \sum_{l=1}^{n-\Delta} p_n(l). \tag{39}$$

C. Maximum-Likelihood Decoding

The following lemma provides an upper bound on $p_n(l)$ for ML decoding with the score function $S_l = p_{X_l^n}(\bar{x}_l^n)$ specified in (32). The proof is given in Appendix A and uses a Chernoff bounding argument similar to [12].

Lemma 1:

$$p_n(l) \leq \exp\{-(n-l+1)E_{pt,x}(R) + 1\},$$

where the form of $E_{pt,x}(R)$ is given in (16).

Using Lemma 1 in (39) gives

$$\begin{aligned}
&\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \\
&\leq \sum_{l=1}^{n-\Delta} \exp\{-(n-l+1)E_{pt,x}(R) + 1\} && (40)\\
&= \sum_{l=1}^{n-\Delta} \exp\{-(n-l+1-\Delta)E_{pt,x}(R)\} \exp\{-\Delta E_{pt,x}(R)+1\} \\
&\leq K_0 \exp\{-\Delta E_{pt,x}(R)\} && (41)
\end{aligned}$$

In (41) we pull out the exponent in Δ. The remaining summation is a geometric sum over decaying exponentials and can thus can be bounded by some constant K_0, into which we've also incorporated the exp{1} scaling, which resulted from non-integer rates. This proves Theorem 2 for ML decoding.

The derivation illustrates the insight that sequential decision made for each symbol is analogous to a classic block-coding problem. This is because we only need to decide between sequences that start to differ in the symbol we are trying to estimate — previous symbols have been fixed, and subsequent symbols are not yet in question. Thus, all sequences that could lead to different estimates of symbol l are binned independently for the remainder of the block. This is why the error exponent we derive equals Gallager's block coding exponent [12]. Since the error exponent for each block-decoding problem is the same, the dominant error event is the hard-decision with the shortest block-length. This corresponds to the last symbol we need to estimate and its block-length equals the estimation delay Δ.

In the remainder of the paper we will assume that all rates are integer. This will greatly simplify notation and, as we have seen above non-integer rates only slightly affect the constant in front of the exponential decay, i.e., the K_0 in (41).

D. Universal Decoding

The following lemma provides an upper bound on $p_n(l)$ for universal decoding with the score function $S_l(\bar{x}) = 1/H(\bar{x}_l^n)$ defined in (33).

$$p_n(l) = \sum_{x^n} \Pr\big[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } H(\tilde{x}_l^n) \le H(x_l^n)\big] p_{\mathbf{x}}(x^n), \quad (42)$$

and the following lemma bounds $p_n(l)$.

Lemma 2: For minimum empirical suffix-entropy decoding, $p_n(l) \le (n-l+2)^{2|\mathcal{X}|} \exp\{-(n-l+1)E_{pt,x}(R)\}$.

Proof: We define P^{n-l} to be the type of length-$(n-l+1)$ sequence x_l^n, and $\mathcal{T}_{P^{n-l}}$ to be the corresponding type class so that $x_l^n \in \mathcal{T}_{P^{n-l}}$. Analogous definitions hold for $\tilde{P}^{n-l}$ and $\tilde{x}_l^n$. We rewrite the constraint $H(\tilde{x}_l^n) \le H(x_l^n)$ as $H(\tilde{P}^{n-l}) \le H(P^{n-l})$. Thus,

$$\begin{aligned}
p_n(l) &= \sum_{x^n} \Pr\big[\exists\, \tilde{x}^n \in \mathcal{B}_x(x^n) \cap \mathcal{F}_n(l, x^n) \text{ s.t. } H(\tilde{x}_l^n) \le H(x_l^n)\big] p_{\mathbf{x}}(x^n) \\
&\le \sum_{x_1^n} \min\Big[1, \sum_{\substack{\tilde{x}_1^n \in \mathcal{F}_n(l,x^n) \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} \Pr[\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)]\Big] p_{\mathbf{x}}(x^n) \\
&= \sum_{x_1^{l-1}, x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} \exp\{-(n-l+1)R\}\Big] p_{\mathbf{x}}(x^{l-1}) p_{\mathbf{x}}(x_l^n) \\
&= \sum_{x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} \exp\{-(n-l+1)R\}\Big] p_{\mathbf{x}}(x_l^n) \quad (43) \\
&= \sum_{P^{n-l}} \sum_{x_l^n \in \mathcal{T}_{P^{n-l}}} \min\Big[1, \sum_{\substack{\tilde{P}^{n-l} \text{ s.t.} \\ H(\tilde{P}^{n-l}) \le H(P^{n-l})}} \sum_{\tilde{x}_l^n \in \mathcal{T}_{\tilde{P}^{n-l}}} \exp\{-(n-l+1)R\}\Big] p_{\mathbf{x}}(x_l^n) \quad (44) \\
&\le \sum_{P^{n-l}} \sum_{x_l^n \in \mathcal{T}_{P^{n-l}}} \min\Big[1, (n-l+2)^{|\mathcal{X}|} \exp\{-(n-l)[R - H(P^{n-l})]\}\Big] p_{\mathbf{x}}(x_l^n) \quad (45) \\
&\le (n-l+2)^{|\mathcal{X}|} \sum_{P^{n-l}} \sum_{x_l^n \in \mathcal{T}_{P^{n-l}}} \exp\{-(n-l+1)[|R - H(P^{n-l})|^+]\} \exp\{-(n-l+1)[D(P^{n-l}\|p_x) + H(P^{n-l})]\} \quad (46) \\
&\le (n-l+2)^{|\mathcal{X}|} \sum_{P^{n-l}} \exp\{-(n-l+1) \inf_q [D(q\|p_x) + |R - H(q)|^+]\} \quad (47) \\
&\le (n-l+2)^{2|\mathcal{X}|} \exp\{-(n-l+1)E_{pt,x}(R)\}. \quad (48)
\end{aligned}$$

In going from (44) to (45) first note that the argument of the inner-most summation (over $\tilde{x}_l^n$) does not depend on $\mathbf{x}$. We then use the following relations: (i) $\sum_{\tilde{x}_l^n \in \mathcal{T}_{\tilde{P}^{n-l}}} = |\mathcal{T}_{\tilde{P}^{n-l}}| \le \exp\{(n-l+1)H(\tilde{P}^{n-l})\}$, which is a standard bound on the size of the type class, (ii) $H(\tilde{P}^{n-l}) \le H(P^{n-l})$ by the minimum-suffix-entropy decoding rule, and (iii) the polynomial bound on the number of types, $|\{\tilde{P}^{n-l}\}| \le (n-l+2)^{|\mathcal{X}|}$. In (46) we recall the function definition $|\cdot|^+ \triangleq \max\{0, \cdot\}$. We pull the polynomial term out of the minimization and use $p_{\mathbf{x}}(x_l^n) = \exp\{-(n-l+1)[D(P^{n-l}\|p_x) + H(P^{n-l})]\}$ for all $p_{\mathbf{x}}(x_l^n) \in \mathcal{T}_{P^{n-l}}$. It is also in (46) that we see why we use a minimum empirical suffix-entropy decoding rule instead of a minimum empirical block-entropy decoding rule. If we had not marginalized out over x^{l-1} in (43) then we would have a polynomial term out front in terms of n rather than $n-l$, which for large n could dominate the exponential decay in $n-l$. As the expression in (47) no longer depends on x_l^n, we simplify by using $|\mathcal{T}_{P^{n-l}}| \le \exp\{(n-l+1)H(P^{n-l})\}$. In (48) we use the form of $E_{pt,x}(\cdot)$ specified in (17) together with the polynomial bound on the number of types. ■

Starting from (39) together the definition of $p_n(l)$ for minimum-suffix decoding from (42) and Lemma 2 provides a bound on the probability of error for universal decoding. Using the definition of $p_n(l)$ in (42) we have

$$\begin{aligned}
\Pr[\hat{x}^{n-\Delta} \ne x^{n-\Delta}] &\le \sum_{l=1}^{n-\Delta} p_n(l) \\
&\le \sum_{l=1}^{n-\Delta} (n-l+2)^{2|\mathcal{X}|} \exp\{-(n-l+1)E_{pt,x}(R)\} \\
&\le \sum_{l=1}^{n-\Delta} K_1 \exp\{-(n-l+1)[E_{pt,x}(R) - \eta]\} \quad (49) \\
&\le K_2 \exp\{-\Delta[E_{pt,x}(R) - \eta]\} \quad (50)
\end{aligned}$$

In (49) we incorporate the polynomial into the exponent, resulting in the constants K_1 and η. Namely, for all $a > 0$, $b > 0$, there exists a C such that $z^a \le C\exp\{b(z-1)\}$ for all $z \ge 1$. We then make explicit the delay-dependent term. Pulling out the exponent in Δ, the remaining summation is a sum over decaying exponentials, and can be bounded by a constant. Together with K_1, this gives the constant K_2 in (50). This proves that universal coding achieves the exponent specified in Theorem 2. Note that the η in (50) does not enter the optimization because $\eta > 0$ can be picked equal to any arbitrarily small constant. The choice of η only effects the constant K in the theorem.

E. Comment on Streaming Source Coding With Side Information at the Decoder

If a random sequence y^n, related to the source x^n through a discrete memoryless channel, is observed at the decoder, then this side information can be used to reduce the rate of the source code. In the model we study $p_{\mathbf{x},\mathbf{y}}(x^n, y^n) = \prod_{i=1}^n p_{x,y}(x_i, y_i) = \prod_{i=1}^n p_{x|y}(x_i|y_i) p_y(y_i)$. The source x^n is observed at the encoder, and the decoder, which observes y^n and a bit stream from the encoder, wants to estimate each source symbol x_i with a probability of error that decreases exponentially in the decoding delay Δ.

The earlier analysis of this section applies to this problem with a few very minor modifications. For ML decoding, we need to pick the sequence with the maximum conditional probability given y^n. The error exponent can be derived using a similar Chernoff bounding argument as in Section V.

For universal decoding, the only change is that we now use a minimum suffix conditional-entropy decoder that compares sequence pairs $(\bar{x}^n, y^n)$ and $(\bar{\tilde{x}}^n, y^n)$. In terms of the analysis, one change enters in (35) where we must also sum over the possible side information sequences. And in (44) the entropy condition in the summation over $\tilde{\mathbf{x}}$ changes to $H(\tilde{x}_{l+1}^n|y_{l+1}^n) < H(x_{l+1}^n|y_{l+1}^n)$ (or the equivalent type notation). Since y^n is observed at the decoder, there is no ambiguity in the side information. Therefore, this condition is equivalent to $H(\tilde{x}_{l+1}^n, y_{l+1}^n) < H(x_{l+1}^n, y_{l+1}^n)$.

We do not include the full derivation as no new ideas are required.

VI. Streaming Slepian-Wolf Source Coding

In this section we prove ML decoding yields the form of the error exponent specified in (29) and that universal decoding yields the form of the exponent specified in (30). The equivalence of the two forms is deferred to Appendix C. As with the proof of Theorem 2 we first develop the common core of the proof that pertains to both ML and universal decoding. The development for more than $l > 2$ sources would essentially be the same, just with more notation and additional minimization parameters $\gamma_1, \gamma_2, \ldots, \gamma_{l-1}$.

A. Sequential Joint Scoring Decoders

We now introduce the class of decoders needed for joint decoding. As in Section V both ML and universal decoders can be cast as a type of decision-directed sequential scoring decoder, where different scoring functions are used in each case. The definition is a bit more involved for joint decoders as all possible pairings between $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ sequences, respectively in $\mathcal{B}_x(\mathbf{x})$ and $\mathcal{B}_y(\mathbf{y})$, must be considered.

Definition 4: A sequential joint scoring decoder constructs its estimate in a sequential manner starting from $l = 1$ where

$$\hat{x}_l = \begin{cases} \bar{x}_l & \text{if for some } \bar{\mathbf{x}} \in \mathcal{B}_x(\mathbf{x}) \text{ s.t. } \bar{x}^{l-1} = \hat{x}^{l-1} \\ & \text{there exists a } \bar{\mathbf{y}} \in \mathcal{B}_y(\mathbf{y}) \text{ s.t.} \\ & \text{for all } (\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in \mathcal{B}_x(\mathbf{x}) \times \mathcal{B}_y(\mathbf{y}) \text{ where} \\ & \tilde{x}^{l-1} = \hat{x}^{l-1}, \tilde{x}_l \neq \bar{x}_l \\ & S_{l,k}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \geq S_{l,k}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \text{ for the } k \in \{1, 2, \ldots, n\} \\ & \text{s.t. } \tilde{y}^{k-1} = \bar{y}^{k-1}, \tilde{y}_k \neq \bar{y}_k \\ ? & \text{otherwise} \end{cases} \quad (51)$$

where $S_{l,k}(\cdot, \cdot)$ is a (possibly time-dependent) joint scoring function and the (failure) symbol "?" is included in case such an $\bar{\mathbf{x}}$ does not exist for some l. Ties are resolved randomly. Just as with the (non-joint) sequential scoring decoders of Definition 3 the estimate of $\mathbf{x}$ is built up sequentially, but now all possible pairings with sequences in $\mathcal{B}_y(\mathbf{y})$ are considered. The "error" symbol "?" is allowed in case there is no $\bar{\mathbf{x}} \in \mathcal{B}_x(\mathbf{x})$ that satisfies the definition. In such an event all subsequent symbol estimates, i.e., $\hat{x}_{l+1}^n$ are also equal to "?", at least until the next parity symbols become available to the decoder.

In the case of known statistics, we use the scoring function

$$S_{l,k}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = p_{\mathbf{x},\mathbf{y}}(\bar{\mathbf{x}}_{\min\{l,k\}}^n, \bar{\mathbf{y}}_{\min\{l,k\}}^n). \quad (52)$$

where, since we only consider i.i.d. sources, the dimension of the subscripts in $p_{\mathbf{x},\mathbf{y}}(\cdot, \cdot)$ can be inferred from the arguments. Just as was the case for (32), this scoring function leads to the ML estimate being constructed in a sequential manner.

In the case of unknown statistics, we use the scoring function

$$S_{l,k}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = 1/H_S(l, k, \bar{\mathbf{x}}, \bar{\mathbf{y}}) \quad (53)$$

where $H_S(\cdot, \cdot, \cdot, \cdot)$ is the "weighted empirical suffix-entropy" function, defined as

$$H_S(l, k, \bar{\mathbf{x}}, \bar{\mathbf{y}}) = \begin{cases} H\left(\bar{x}_l^n, \bar{y}_l^n\right) & \text{if } l = k \\ \frac{k-l}{n+1-l} H\left(\bar{x}_l^{k-1} \middle| \bar{y}_l^{k-1}\right) + \frac{n+1-k}{n+1-l} H\left(\bar{x}_k^n, \bar{y}_k^n\right) & \text{if } l < k \\ \frac{l-k}{n+1-k} H\left(\bar{y}_k^{l-1} \middle| \bar{x}_k^{l-1}\right) + \frac{n+1-l}{n+1-k} H\left(\bar{x}_l^n, \bar{y}_l^n\right) & \text{if } l > k. \end{cases} \quad (54)$$

Due to the fact that $H_S(l, k, \bar{\mathbf{x}}, \bar{\mathbf{y}})$ weights the empirical suffix entropies differently, based upon the values of l and k, we term the resulting decoder the "minimum weighted empirical suffix entropy" decoder.

Note that the form of the two scoring functions (52) and (54) is more similar than may initially appear. For instance compare the ML scores of two pairs $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ and $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ where $\bar{x}^{l-1} = \tilde{x}^{l-1}$ and $\bar{y}^{k-1} = \tilde{y}^{k-1}$, but $\bar{x}_l \neq \tilde{x}_l$ and $\bar{y}_k \neq \tilde{y}_k$, and $l < k$. Then the question of whether $S_{l,k}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is larger than $S_{l,k}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ is the same as asking whether $-\log p_{\mathbf{x}|\mathbf{y}}(\bar{x}_l^{k-1}|b^{k-l}) - \log p_{\mathbf{x},\mathbf{y}}(\bar{x}_k^n, \bar{y}_k^n)$ is smaller than $-\log p_{\mathbf{x}|\mathbf{y}}(\tilde{x}_l^{k-1}|b^{k-l}) - \log p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n, \tilde{y}_k^n)$ where $b^{k-1} = \bar{y}_l^{k-1} = \tilde{y}_l^{k-1}$. The analog to the weightings of (54) comes from the dimensions of the various subsequences.

An error can only occur if there is some erroneous source pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in \mathcal{B}_x(\mathbf{x}) \times \mathcal{B}_y(\mathbf{y})$ such that $S_{l,k}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq S_{l,k}(\mathbf{x}, \mathbf{y})$ for some $l \leq n - \Delta$. Otherwise, the realized source $\mathbf{x}$ will match $\hat{x}^n$ at least through the $n - \Delta$th symbol. For both our choices of score functions we show in the following sections that the probability of such an event decays exponentially in Δ.

B. Error Analysis of Sequential Joint Scoring Decoders

We follow the same approach to prove Theorem 6 that we used for lossless streaming source coding in Section V. We first develop the common core of the proof for sequential joint scoring decoders in general. We then specialize the scoring function to the ML and universal scoring functions, (52) and (54), respectively.

In Theorem 6 three error events are considered: (a) $\Pr[x^{n-\Delta} \neq \hat{x}^{n-\Delta}]$, (b) $\Pr[y^{n-\Delta} \neq \hat{y}^{n-\Delta}]$, and (c) $\Pr[(x^{n-\Delta}, y^{n-\Delta}) \neq (\hat{x}^{n-\Delta}, \hat{y}^{n-\Delta})]$. We develop the error exponent for event (a). The exponent of event (b) follows from a similar derivation, and that of event (c) from an application of the union bound resulting in an exponent that is the minimum of the exponents of events (a) and (b).

For there to be a decoding error there must be some spurious source pair $(\tilde{x}^n, \tilde{y}^n)$ that satisfies three conditions: (i) $\tilde{x}^n \in \mathcal{B}_x(x^n)$ and $\tilde{y}^n \in \mathcal{B}_y(y^n)$, (ii) $\tilde{x}_l \neq x_l$ for some $l \leq n - \Delta$ while $\bar{x}^{l-1} = x^{l-1}$ and (iii) for the time index l of event (ii) and for the $k \in \{1, \ldots, n\}$ such that $\tilde{y}^{k-1} = y^{k-1}$ but $\tilde{y}_k \neq y_k$, the spurious source pair has a higher score than the true pair, i.e., $S_{l,k}(\tilde{x}^n, \tilde{y}^n) > S_{l,k}(x^n, y^n)$.

As in (34) we again introduce a partition of source sequences to track condition (ii). This time we partition all source pairs $(\tilde{x}^n, \tilde{y}^n) \in \{\mathcal{X}^n, \mathcal{Y}^n\}$ into sets $\mathcal{F}_n(l, k, x^n, y^n)$ defined by the times l and k at which $\tilde{x}^n$ and $\tilde{y}^n$ respectively diverge from the realized source sequences. Formally,

$$\mathcal{F}_n(l,k,x^n,y^n) = \{(\bar{x}^n,\bar{y}^n) \in \mathcal{X}^n \times \mathcal{Y}^n \text{ s.t. } \bar{x}^{l-1} = x^{l-1}, \bar{x}_l \neq x_l, \bar{y}^{k-1} = y^{k-1}, \bar{y}_k \neq y_k\}, \quad (55)$$

and $\mathcal{F}_n(n+1, n+1, x^n, y^n) = \{(x^n, y^n)\}$ so $\cup_{l=1}^{n+1} \cup_{k=1}^{n+1} \mathcal{F}_n(l, k, x^n, y^n) = \mathcal{X}^n \times \mathcal{Y}^n$. In contrast to streaming point-to-point or side-information coding (cf. (55) with (34)), the partition is now doubly-indexed. To find the dominant error event, we will need to search over both indices. This search is the reason why the streaming exponents differ from the block coding exponents, manifesting itself in the γ parameter of Theorem 6.

We now bound the marginal error probability $\Pr[x^{n-\Delta} \neq \hat{x}^{n-\Delta}]$.

$$\begin{aligned}
&\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \\
&= \sum_{x^n,y^n} \Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta} | x^n = x^n, y^n = y^n] p_{\mathbf{x},\mathbf{y}}(x^n, y^n) \\
&= \sum_{x^n,y^n} p_{\mathbf{x},\mathbf{y}}(x^n, y^n) \Big\{ \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} \\
&\qquad \Pr\left[\exists\, (\tilde{x}^n, \tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l,k,x^n,y^n) \right. \\
&\qquad\qquad \left. \text{s.t. } S_{l,k}(\tilde{x}^n, \tilde{y}^n) \geq S_{l,k}(x^n, y^n)\right]\Big\} \qquad (56) \\
&= \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} \Big\{ \sum_{x^n,y^n} p_{\mathbf{x},\mathbf{y}}(x^n, y^n) \\
&\qquad \Pr\left[\exists\, (\tilde{x}^n, \tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l,k,x^n,y^n) \right. \\
&\qquad\qquad \left. \text{s.t. } S_{l,k}(\tilde{x}^n, \tilde{y}^n) \geq S_{l,k}(x^n, y^n)\right]\Big\} \qquad (57)
\end{aligned}$$

where in (56) we decompose the error event according to conditions (i)–(iii) discussed above, and the equality results from the fact that $\mathcal{F}_n(l, k, x^n, y^n) \cap \mathcal{F}_n(l', k', x^n, y^n) = \{\}$, the null set, for $(l, k) \neq (l', k')$. Defining $p_n(l, k)$ as

$$\begin{aligned}
p_n(l,k) &= \sum_{x^n,y^n} p_{\mathbf{x},\mathbf{y}}(x^n, y^n) \\
&\quad \Pr\left[\exists\, (\tilde{x}^n, \tilde{y}^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l,k,x^n,y^n) \right. \\
&\qquad \left. \text{s.t. } S_{l,k}(\tilde{x}^n, \tilde{y}^n) \geq S_{l,k}(x^n, y^n)\right]. \qquad (58)
\end{aligned}$$

and substituting the definition into (57) we get

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] = \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} p_n(l,k). \qquad (59)$$

C. Maximum Likelihood Decoding

To develop our results for ML decoding we use use the joint score function of (52) in (58). With this choice the following lemma, proved in Appendix B, provides an upper bound on $p_n(l, k)$.

Lemma 3:

$$\begin{aligned}
p_n(l,k) &\leq \exp\left\{-(n-l+1) E_x\left(R_x, R_y, \frac{k-l}{n-l+1}\right)\right\} \quad \text{if } l \leq k, \\
p_n(l,k) &\leq \exp\left\{-(n-k+1) E_y\left(R_x, R_y, \frac{l-k}{n-k+1}\right)\right\} \quad \text{if } l \geq k,
\end{aligned} \qquad (60)$$

where $E_x(R_x, R_y, \gamma)$ and $E_y(R_x, R_y, \gamma)$ are defined in (29). Notice that $l, k \leq n$ and that for $l \leq k$ the fraction $\frac{k-l}{n-l+1} \in [0, 1]$ serves as γ in the error exponent $E_x(R_x, R_y, \gamma)$. An analogous discussion holds for $l \geq k$.

We use Lemma 3 together with (59) to bound $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$ for two distinct cases. The first, simpler case, is when $\inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma) > \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma)$. To bound $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$ in this case, we split the sum over the $p_n(l, k)$ into two terms, as is visualized in Fig. 9. There are $(n + 1) \times (n - \Delta)$ such events to account for. In Fig. 9 these are inside the box. The probability of the event within each oval are summed together to give an upper bound on $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$. We add extra probabilities outside of the box but within the ovals to make the summation symmetric thus simpler. Those extra error events do not impact the error exponent because this case assumes that $\inf_{\gamma\in[0,1]} E_y(R_x, R_y, \rho, \gamma) \geq \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \rho, \gamma)$. The possible dominant error events are highlighted in Fig. 9. Thus,

$$\begin{aligned}
\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] &= \sum_{l=1}^{n-\Delta} \sum_{k=l}^{n+1} p_n(l,k) + \sum_{k=1}^{n-\Delta} \sum_{l=k}^{n+1} p_n(l,k) \qquad (61) \\
&\leq \sum_{l=1}^{n-\Delta} \sum_{k=l}^{n+1} \exp\{-(n-l+1) \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma)\} \\
&\quad + \sum_{k=1}^{n-\Delta} \sum_{l=k}^{n+1} \exp\{-(n-k+1) \inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma)\} \qquad (62) \\
&= \sum_{l=1}^{n-\Delta} \Big[(n-l+2) \exp\{-(n-l+1) \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma)\} \\
&\quad + \sum_{k=1}^{n-\Delta} \Big[(n-k+2) \exp\{-(n-k+1) \inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma)\}\Big] \\
&\leq 2 \sum_{l=1}^{n-\Delta} \Big[(n-l+2) \exp\{-(n-l+1) \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma)\}\Big] \qquad (63) \\
&\leq \sum_{l=1}^{n-\Delta} C_1 \exp\{-(n-l+2)[\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma) - \alpha]\} \qquad (64) \\
&\leq C_2 \exp\{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma) - \alpha]\} \qquad (65)
\end{aligned}$$

Equation (61) follows directly from (59), in the first term $l \leq k$, in the second term $l \geq k$. In (62), we use Lemma 3. In (63) we use the assumption that $\inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma) > \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma)$. In (64) the $\alpha > 0$ results from incorporating the polynomial into the first exponent, and can be chosen as small as desired. Combining terms and summing out the decaying exponential yield the bound (65).

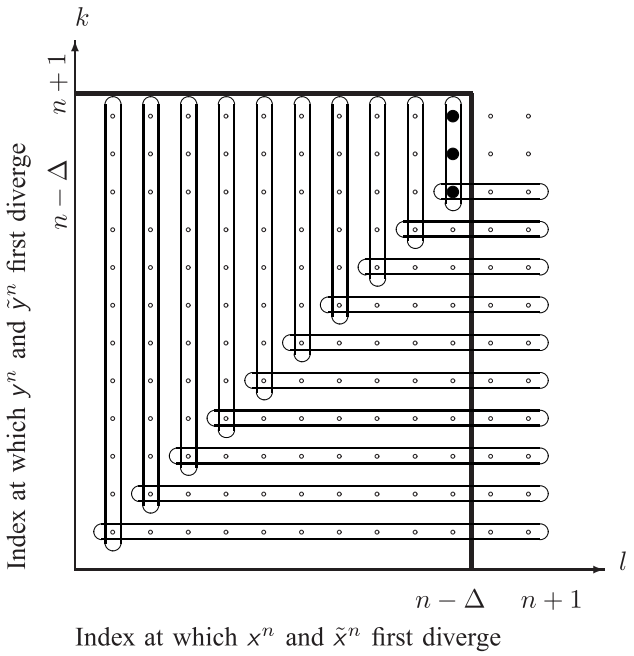

Fig. 9. Two dimensional plot of the error probabilities $p_n(l, k)$, corresponding to error events (l, k), contributing to $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$ in the situation where $\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \rho, \gamma) \geq \inf_{\gamma \in [0,1]} E_x(R_x, R_y, \rho, \gamma)$.

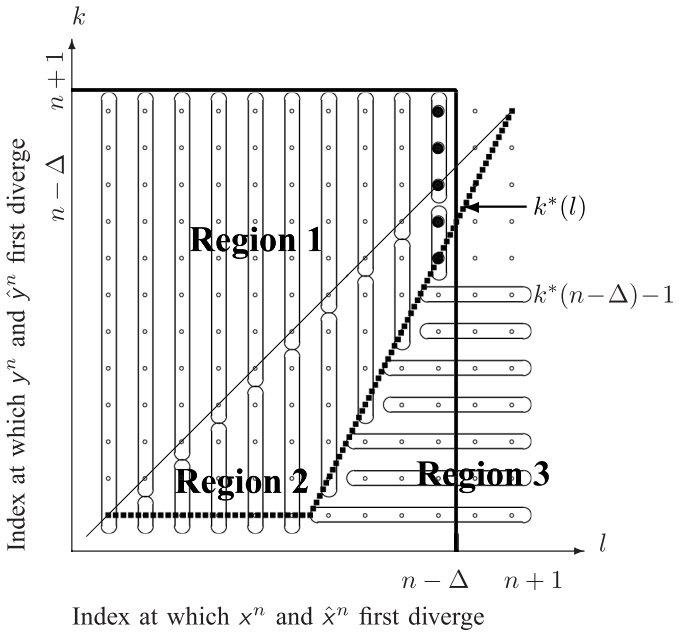

Fig. 10. Two dimensional plot of the error probabilities $p_n(l, k)$, corresponding to error events (l, k), contributing to $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$ in the situation where $\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \gamma) < \inf_{\gamma \in [0,1]} E_x(R_x, R_y, \gamma)$.

The second, more involved case, is when $\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \rho, \gamma) < \inf_{\gamma \in [0,1]} E_x(R_x, R_y, \rho, \gamma)$. To bound $\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$, we could use the same bounding technique used in the first case. This gives the error exponent $\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \gamma)$ which is generally smaller than what we can get by dividing the error events in a new grouping shown in Fig. 10. In this situation we split (59) into three terms, as visualized in Fig. 10. Just as in the first case shown in Fig. 9, there are $(n+1) \times (n-\Delta)$ such events to account for (those inside the box). The error events are partitioned into 3 regions. Region 2 and 3 are separated by $k^*(l)$ using a dotted line. In region 3, we add extra probabilities outside of the box but within the ovals to make the summation simpler. Those extra error events do not affect the error exponent as shown in the proof. The possible dominant error events are highlighted in Fig. 10. Thus,

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}] \leq \sum_{l=1}^{n-\Delta} \sum_{k=l}^{n+1} p_n(l,k) + \sum_{l=1}^{n-\Delta} \sum_{k=k^*(l)}^{l-1} p_n(l,k) + \sum_{l=1}^{n-\Delta} \sum_{k=1}^{k^*(l)-1} p_n(l,k) \tag{66}$$

Where $\sum_{k=1}^{0} p_k = 0$. The lower boundary of Region 2 is $k^*(l) \geq 1$ as a function of n and l:

$$k^*(l) = \max\left\{1, n+1-\left\lceil \frac{\inf_{\gamma \in [0,1]} E_x(R_x, R_y, \gamma)}{\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \gamma)} \right\rceil (n+1-l)\right\} \tag{67}$$

For compactness in the ensuing development we use G (always non-negative) to denote the ceiling of the ratio of exponents, i.e.,

$$G = \left\lceil \frac{\inf_{\gamma \in [0,1]} E_x(R_x, R_y, \gamma)}{\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \gamma)} \right\rceil. \tag{68}$$

Note that when $\inf_{\gamma \in [0,1]} E_y(R_x, R_y, \gamma)$ is greater than $\inf_{\gamma \in [0,1]} E_x(R_x, R_y, \gamma)$ then $G = 1$ and region two of Fig. 10 disappears. In other words, the middle term of (66) equals zero. This was the first case considered. We now consider the situation in which $G \geq 2$.

The first term of (66), i.e., region one in Fig. 10 where $l \leq k$, is bounded in the same way that the first term of (61) is, giving

$$\sum_{l=1}^{n-\Delta} \sum_{k=l}^{n+1} p_n(l,k) \leq C_2 \exp\left\{-\Delta\left[\inf_{\gamma \in [0,1]} E_x(R_x, R_y, \gamma) - \alpha\right]\right\}. \tag{69}$$

In Fig. 10, region two is upper bounded by the 45-degree line, and lower bounded by $k^*(l)$. The second term of (66), corresponding to this region where $l \geq k$,

$$\begin{aligned} &\sum_{l=1}^{n-\Delta} \sum_{k=k^*(l)}^{l-1} p_n(l,k) \\ &\leq \sum_{l=1}^{n-\Delta} \sum_{k=k^*(l)}^{l-1} \exp\left\{-(n-k+1) E_y\left(R_x, R_y, \frac{l-k}{n-k+1}\right)\right\} \\ &= \sum_{l=1}^{n-\Delta} \sum_{k=k^*(l)}^{l-1} \exp\left\{-(n-l+1)\frac{n-k+1}{n-l+1} E_y\left(R_x, R_y, \frac{l-k}{n-k+1}\right)\right\} \end{aligned} \tag{70}$$

$$\leq \sum_{l=1}^{n-\Delta} \sum_{k=k^*(l)}^{l-1} \exp\left\{-(n-l+1) \inf_{\gamma \in [0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma)\right\} \tag{71}$$

$$= \sum_{l=1}^{n-\Delta} (l-k^*(l)) \exp\left\{-(n-l+1) \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma)\right\} \tag{72}$$

In (70) we note that $l \geq k$, so define $\frac{l-k}{n-k+1} = \gamma$ as in (71). Then $\frac{n-k+1}{n-l+1} = \frac{1}{1-\gamma}$.

The third term of (66), i.e., the intersection of region three and the "box" in Fig. 10 where $l \geq k$, can be bounded as,

$$\sum_{l=1}^{n-\Delta} \sum_{k=1}^{k^*(l)-1} p_n(l,k)$$

$$\leq \sum_{l=1}^{n+1} \sum_{k=1}^{\min\{l,k^*(n-\Delta)-1\}} p_n(l,k) \tag{73}$$

$$= \sum_{k=1}^{k^*(n-\Delta)-1} \sum_{l=k}^{n+1} p_n(l,k) \tag{74}$$

$$\leq \sum_{k=1}^{k^*(n-\Delta)-1} \sum_{l=k}^{n+1} \exp\{-(n-k+1)E_y(R_x, R_y, \frac{l-k}{n-k+1})\}$$

$$\leq \sum_{k=1}^{k^*(n-\Delta)-1} \sum_{l=k}^{n+1} \exp\{-(n-k+1) \inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma)\}$$

$$\leq \sum_{k=1}^{k^*(n-\Delta)-1} (n-k+2) \exp\{-(n-k+1) \inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma)\} \tag{75}$$

In (73) we note that $l \leq n-\Delta$ thus $k^*(n-\Delta)-1 \geq k^*(l)-1$, also $l \geq 1$, so $l \geq k^*(l)-1$. This can be visualized in Fig. 10 as we extend the summation from the intersection of the "box" and region 3 to the whole region under the diagonal line and the horizontal line $k = k^*(n-\Delta)-1$. In (74) we simply switch the order of the summation.

Finally when $G \geq 2$, we substitute (69), (72), and (75) into (66) to give

$$\Pr[\hat{x}^{n-\Delta} \neq x^{n-\Delta}]$$

$$\leq C_2 \exp\left\{-\Delta\left[\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma) - \alpha\right]\right\}$$

$$+ \sum_{l=1}^{n-\Delta} (l-k^*(l)) \exp\left\{-(n-l+1) \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma)\right\} \tag{76}$$

$$+ \sum_{k=1}^{k^*(n-\Delta)-1} (n-k+2) \exp\left\{-(n-k+1) \inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma)\right\}$$

$$\leq C_2 \exp\left\{-\Delta\left[\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma) - \alpha\right]\right\}$$

$$+ \sum_{l=1}^{n-\Delta} (l-n-1+G(n+1-l))$$

$$\times \exp\left\{-(n-l+1) \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma)\right\}$$

$$+ \sum_{k=1}^{n+1-G(\Delta+1)} (n-k+2) \exp\left\{-(n-k+1) \inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma)\right\} \tag{77}$$

$$\leq C_2 \exp\left\{-\Delta\left[\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma) - \alpha\right]\right\}$$

$$+ (G-1)C_3 \exp\left\{-\Delta\left[\inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma) - \alpha\right]\right\}$$

$$+ C_4 \exp\left\{-\left[\Delta G \inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma) - \alpha\right]\right\}$$

$$\leq C_5 \exp\left\{-\Delta\left[\min\left\{\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma)\right\} - \alpha\right]\right\}. \tag{78}$$

To get (77), we use the fact that $k^*(l) \geq n+1-G(n+1-l)$ from the definition of $k^*(l)$ in (67) to upper bound the second term. We exploit the definition of G to convert the exponent in the third term to $\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma)$. Finally, to get (78) we gather the constants together, sum out over the decaying exponentials, and are limited by the smaller of the two exponents.

One might note that in this proof we regularly double count the error events and add some extra small probabilities to simplify sums. The error exponent is not modified by these manipulations.

D. Universal Decoding

To develop our universal results we use the joint universal scoring function $S_{l,k}(\tilde{x}^n, \tilde{y}^n) = 1/H_S(l,k,\tilde{x}^n, \tilde{y}^n)$ from (54) in (58).

$$p_n(l,k) = \sum_{x^n} \sum_{y^n} p_{\mathbf{xy}}(x^n, y^n)$$

$$\Pr\Big[\exists\, (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x^n) \times \mathcal{B}_y(y^n) \cap \mathcal{F}_n(l,k,x^n,y^n)$$

$$\text{s.t. } H_S(l,k,\tilde{x}^n,\tilde{y}^n) \leq H_S(l,k,x^n,y^n)\Big] \tag{79}$$

The following lemma bound the contributions of each $p_n(l,k)$ to the overall error probability.

Lemma 4: Upper bound on $p_n(l,k)$ for $l \leq k$. For all $\eta > 0$, there exists a constant $K_1 < \infty$, s.t.

$$p_n(l,k) \leq K_1 \exp\{-(n-l+1)[E_x(R_x, R_y, \lambda) - \eta]\}$$

where $\lambda = (k-l)/(n-l+1) \in [0,1]$.

Proof: Starting from (79) we have the steps shown in (80)–(81). In (81) we enumerate all the source sequences in a way that allows us to focus on the types of the important subsequences. We enumerate the possibly misleading candidate sequences in terms of their suffixes types. We restrict the sum to those pairs $(\tilde{x}^n, \tilde{y}^n)$ that could lead to mistaken decoding, defining the compact notation $S(P^{n-k}, P^{k-l}, V^{n-k}, V^{k-l}) \triangleq (k-l)H(V^{k-l}|P^{k-l}) + (n-k+1)H(P^{n-k} \times V^{n-k})$, which is the weighted empirical suffix entropy condition rewritten in terms of types.

Note that the summations within the minimization in (81) do not depend on the arguments within these sums. Thus, we can bound this sum separately to get a bound on the number of possibly misleading source pairs $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$. We bound this sum starting in (82).

In (83) we sum over all $\tilde{x}_l^{k-1} \in \mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})$. In (84) we use standard bounds, e.g., $|\mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})| \leq \exp\{(k-l)$

$$p_n(l,k) \le \sum_{x^n,y^n} \min\Bigg[1, \sum_{\substack{(\tilde{x}^n,\tilde{y}^n)\in\mathcal{F}_n(l,k,x^n,y^n)\\ H_S(l,k,\tilde{x}^n,\tilde{y}^n)\le H_S(l,k,x^n,y^n)}} \Pr[\tilde{x}^n\in\mathcal{B}_x(x^n),\tilde{y}^n\in\mathcal{B}_y(y^n)]\Bigg] p_{\mathbf{x},\mathbf{y}}(x^n,y^n) \tag{80}$$

$$\le \sum_{x_l^n,y_l^n} \min\Bigg[1, \sum_{\substack{(\tilde{x}_l^n,\tilde{y}_l^n)\ \text{s.t.}\ \tilde{y}^{k-1}=y^{k-1}\\ H_S(\tilde{x}_l^n,\tilde{y}_l^n)\le H_S(x_l^n,y_l^n)}} \exp\{-(n-l+1)R_x-(n-k+1)R_y\}\Bigg] p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)$$

$$= \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \sum_{\substack{y_l^{k-1}\in\mathcal{T}_{P^{k-l}},\\ y_k^n\in\mathcal{T}_{P^{n-k}}}} \sum_{\substack{x_l^{k-1}\in\mathcal{T}_{V^{k-l}}(y_l^{k-1}),\\ x_k^n\in\mathcal{T}_{V^{n-k}}(y_k^n)}} \min\Bigg[1, \sum_{\substack{\tilde{P}^{n-k},\tilde{P}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},\tilde{P}^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}}$$

$$\sum_{\tilde{y}_k^n\in\mathcal{T}_{\tilde{P}^{n-k}}} \sum_{\tilde{x}_l^{k-1}\in\mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})} \sum_{\tilde{x}_k^n\in\mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)} \exp\{-(n-l+1)R_x-(n-k+1)R_y\}\Bigg] p_{\mathbf{xy}}(x^n,y^n) \tag{81}$$

$H(\tilde{V}^{k-l}|P^{k-l})\}$ since $y_l^{k-1}\in\mathcal{T}_{P^{k-l}}$. We also sum over all $\tilde{x}_k^n\in\mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)$ and over all $\tilde{y}_k^n\in\mathcal{T}_{\tilde{P}^{n-k}}$ in (84). By definition of the decoding rule $(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ can only lead to a decoding error if $(k-l)H(\tilde{V}^{k-l}|P^{k-l})]+(n-k+1)H(\tilde{P}^{n-k}\times\tilde{V}^{n-k}) < (k-l)H(V^{k-l}|P^{k-l})+(n-k+1)H(P^{n-k}\times V^{n-k})$. In (87) we apply the polynomial bound on the number of types.

We substitute (87) into (81) and pull out the polynomial term, giving (88). In (89) we use the memoryless property of the source, and exponential bounds on the probability of observing (x_l^{k-1},y_l^{k-1}) and (x_k^n,y_k^n). In (90) we pull out $(n-l+1)$ from all terms, noticing that $\lambda=(k-l)/(n-l+1)\in[0,1]$ and $\bar{\lambda}\triangleq 1-\lambda=(n-k+1)/(n-l+1)$. In (91) we minimize the exponent over all choices of distributions $p_{\tilde{x},\tilde{y}}$ and $p_{\bar{x},\bar{y}}$. In (92) we define the universal random coding exponent $E_x(R_x,R_y,\lambda)\triangleq\inf_{\tilde{x},\tilde{y},\bar{x},\bar{y}}\{\lambda D(p_{\tilde{x},\tilde{y}}\|p_{x,y})+\bar{\lambda}D(p_{\bar{x},\bar{y}}\|p_{x,y})+\left|\lambda[R_x-H(\tilde{x}|\tilde{y})]+\bar{\lambda}[R_x+R_y-H(\bar{x},\bar{y})]\right|^+\}$ where $0\le\lambda\le 1$ and $\bar{\lambda}=1-\lambda$. We also incorporate the number of conditional and marginal types into the polynomial bound, as well as the sum over k, and then push the polynomial into the exponent since for any polynomial F, $\forall E,\epsilon>0$, there exists $C>0$, s.t. $F(\Delta)e^{-\Delta E}\le Ce^{-\Delta(E-\epsilon)}$. ■

A similar derivation yields a bound on $p_n(l,k)$ for $l\ge k$.

Using Lemma 4 in (59) and following the same steps as in the derivation for ML decoding of Section VI-C yields (30).

VII. Future Directions

Stationary-Ergodic Sources and Universality: In [3] the block-coding proofs of the Slepian-Wolf problem are extended to stationary-ergodic sources using AEP arguments. To have a similar extension to the streaming context it is possible that additional regularity conditions will be required so that error exponents can be achieved. To additionally achieve universality over non-memoryless sources further technical restrictions will be required. For the specific case of distributed Markov sources however, all the arguments in this paper should generalize easily by following an approach similar to that taken in [24]: the source can be "segmented" into small blocks and the endpoints (for a Markov source of known order k, the endpoint is just k successive symbols at the end of the block) of the blocks can be encoded perfectly at an arbitrarily small rate by making the "small" blocks long enough. Conditioned on these endpoints, the blocks are then i.i.d. with the endpoints representing a third stream of perfectly known side-information.

Upper Bounds and Demonstrating Optimal Delays: This paper focused on achievable exponents for the two-encoder case and presented upper bounds for the side-information at the decoder case. The upper bounds followed by extending single encoder arguments from [22] and do not immediately generalize to the case of multiple encoders.

Trading Off Error Exponents for the Different Source Terminals: For multiple terminal systems, different error exponents can be achieved for different users or different sources. For channel coding, the encoders can choose different distributions while generating the randomized code book to achieve an error exponent trade-off among different users. In [30], the error exponent region is studied for the Gaussian multiple access channel and the broadcast channel within the block-coding paradigm. It is unclear whether similar trade offs are possible within the streaming Slepian Wolf problems considered here since there is nothing immediately comparable to the flexibility we have in choosing the "input distribution" for channel coding problems.

Adaptation and Limited Feedback: A final interesting extension is to adaptive universal streaming Slepian Wolf encoders. The decoders we use in this paper are based on empirical statistics. Therefore they can be used even if source statistics are unknown. The current proposal will work regardless of source and side information statistics as long as the conditional entropy $H(x|y)$ is less than the encoding rate. Even if there is uncertainty in statistics, the sequential nature of the coding system would enable the system to adapt on-line to the unknown entropy rate if some feedback channel is available. The feedback channel would be used to order increases (or decreases) in the binning rate. An increase (or decrease) could be triggered by examining the difference between two quantities: the minimal empirical joint entropy between the decoded sequence and observation, and the empirical joint entropy between the particular sequence and observation yielding the second-lowest joint entropy. If there is a large difference between these two entropies, we are using rate excessively, and the rate of communication can be reduced. If the difference is negligible, then it's likely we are not decoding correctly. Our target should be to keep this difference at roughly ϵ. In the current context, this is analogous to the rate margin by which we choose to exceed the known conditional entropy.

$$\sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \sum_{\tilde{y}_k^n\in\mathcal{T}_{\tilde{P}^{n-k}}} \sum_{\tilde{x}_l^{k-1}\in\mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})} \sum_{\tilde{x}_k^n\in\mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)} \tag{82}$$

$$\le \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \sum_{\tilde{y}_k^n\in\mathcal{T}_{\tilde{P}^{n-k}}} |\mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})||\mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)| \tag{83}$$

$$\le \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} |\mathcal{T}_{\tilde{P}^{n-k}}|\exp\{(k-l)H(\tilde{V}^{k-l}|P^{k-l})\}\exp\{(n-k+1)H(\tilde{V}^{n-k}|\tilde{P}^{n-k})\} \tag{84}$$

$$\le \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \exp\{(k-l)H(\tilde{V}^{k-l}|P^{k-l})+(n-k+1)H(\tilde{P}^{n-k}\times\tilde{V}^{n-k})\} \tag{85}$$

$$\le \sum_{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}} \exp\{(k-l)H(V^{k-l}|P^{k-l})+(n-k+1)H(P^{n-k}\times V^{n-k})\} \tag{86}$$

$$\le (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|}\exp\{(k-l)H(V^{k-l}|P^{k-l})+(n-k+1)H(P^{n-k}\times V^{n-k})\} \tag{87}$$

$$p_n(l,k) \le (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \sum_{\substack{y_l^{k-1}\in\mathcal{T}_{P^{k-l}},\\ y_k^n\in\mathcal{T}_{P^{n-k}}}} \sum_{\substack{x_l^{k-1}\in\mathcal{T}_{V^{k-l}}(y_l^{k-1}),\\ x_k^n\in\mathcal{T}_{V^{n-k}}(y_k^n)}}$$
$$\min\Big[1,\exp\{-(k-l)[R_x-H(V^{k-l}|P^{k-l})]-(n-k+1)[R_x+R_y-H(V^{n-k}\times P^{n-k})]\}\Big]\, p_{x_l^n,y_l^n}(x_l^n,y_l^n) \tag{88}$$

$$\le (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}}$$
$$\exp\Big\{\max\Big[0,-(k-l)[R_x-H(V^{k-l}|P^{k-l})]-(n-k+1)[R_x+R_y-H(V^{n-k}\times P^{n-k})]\Big]\Big\}$$
$$\exp\Big\{-(k-l)D(V^{k-l}\times P^{k-l}\|p_{x,y})-(n-k+1)D(V^{n-k}\times P^{n-k}\|p_{x,y})\Big\} \tag{89}$$

$$\le (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \exp\Big\{-(n-l+1)\Big[\lambda D(V^{k-l}\times P^{k-l}\|p_{x,y})+\bar{\lambda}D(V^{n-k}\times P^{n-k}\|p_{x,y})$$
$$+\Big|\lambda[R_x-H(V^{k-l}|P^{k-l})]+\bar{\lambda}[R_x+R_y-H(V^{n-k}\times P^{n-k})]\Big|^+\Big]\Big\} \tag{90}$$

$$\le (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \exp\Big\{-(n-l+1)\inf_{\tilde{x},\tilde{y},\bar{x},\bar{y}}\Big[\lambda D(p_{\tilde{x},\tilde{y}}\|p_{x,y})+\bar{\lambda}D(p_{\bar{x},\bar{y}}\|p_{x,y})$$
$$+\big|\lambda[R_x-H(\tilde{x}|\tilde{y})]+\bar{\lambda}[R_x+R_y-H(\bar{x},\bar{y})]\big|^+\Big]\Big\} \tag{91}$$

$$\le (n-l+2)^{4|\mathcal{X}||\mathcal{Y}|}\exp\{-(n-l+1)E_x(R_x,R_y,\lambda)\} \le K_1\exp\{-(n-l+1)[E_x(R_x,R_y,\lambda)-\eta]\} \tag{92}$$

APPENDIX

A. Proof of Lemma 1

In this section we provide the proof of Lemma 1.

$$p_n(l)=\sum_{x^n}\Pr\Big[\exists\,\tilde{x}^n\in\mathcal{B}_x(x^n)\cap\mathcal{F}_n(l,x^n) \text{ s.t. } p_{\mathbf{x}}(\tilde{x}_l^n)\ge p_{\mathbf{x}}(x_l^n)\Big]p_{\mathbf{x}}(x^n)$$

$$\le \sum_{x^n}\min\left[1,\sum_{\substack{\tilde{x}^n\in\mathcal{F}_n(l,x^n)\text{s.t.}\\ p_{\mathbf{x}}(x_l^n)\le p_{\mathbf{x}}(\tilde{x}_l^n)}}\Pr[\tilde{x}^n\in\mathcal{B}_x(x^n)]\right]p_{\mathbf{x}}(x^n) \tag{93}$$

$$=\sum_{x^{l-1},x_l^n}\min\left[1,\sum_{\substack{\tilde{x}_l^n\ \text{s.t.}\\ p_{\mathbf{x}}(x_l^n)<p_{\mathbf{x}}(\tilde{x}_l^n)}}\exp\{-(n-l+1)R+1\}\right]p_{\mathbf{x}}(x^n) \tag{94}$$

$$=\sum_{x_l^n}\min\Big[1,\sum_{\substack{\tilde{x}_l^n\ \text{s.t.}\\ p_{\mathbf{x}}(x_l^n)<p_{\mathbf{x}}(\tilde{x}_l^n)}}\exp\{-(n-l+1)R+1\}\Big]p_{\mathbf{x}}(x_l^n)$$
$$=\sum_{x_l^n}\min\left[1,\sum_{\tilde{x}_l^n}I\left[\frac{p_x(\tilde{x}_l^n)}{p_{\mathbf{x}}(x_l^n)}>1\right]\exp\{-(n-l+1)R+1\}\right]p_{\mathbf{x}}(x_l^n) \tag{95}$$

$$\le\sum_{x_l^n}\min\left[1,\sum_{\tilde{x}_l^n}\min\left[1,\frac{p_x(\tilde{x}_l^n)}{p_x(x_l^n)}\right]\exp\{-(n-l+1)R+1\}\right]p_{\mathbf{x}}(x_l^n)$$

$$\leq \sum_{x_l^n}\left[\sum_{\tilde{x}_l^n}\left[\frac{p_X(\tilde{x}_l^n)}{p_X(x_l^n)}\right]^{\frac{1}{1+\rho}}\exp\{-(n-l+1)R+1\}\right]^{\rho} p_{\mathbf{x}}(x_l^n) \tag{96}$$

$$=\sum_{x_l^n} p_{\mathbf{x}}(x_l^n)^{\frac{1}{1+\rho}}\left[\sum_{\tilde{x}_l^n}\left[p_X(\tilde{x}_l^n)\right]^{\frac{1}{1+\rho}}\right]^{\rho}\exp\{-(n-l+1)\rho R+\rho\}$$

$$=\left[\sum_x p_X(x)^{\frac{1}{1+\rho}}\right]^{(n-l+1)}\left[\sum_x p_X(x)^{\frac{1}{1+\rho}}\right]^{(n-l+1)\rho}\exp\{-(n-l+1)\rho R+\rho\} \tag{97}$$

$$=\left[\sum_x p_X(x)^{\frac{1}{1+\rho}}\right]^{(n-l+1)(1+\rho)}\exp\{-(n-l+1)\rho R+1\}$$

$$=\exp\left\{-(n-l+1)\left[\rho R-(1+\rho)\ln\left(\sum_x p_X(x)^{\frac{1}{1+\rho}}\right)\right]+1\right\}. \tag{98}$$

In (93) the union bound is applied. In (94) we use the fact that after the first symbol in which two sequences differ, the remaining parity bits are independent. The number of such symbols is $\lfloor nR\rfloor - \lfloor (l-1)R\rfloor \geq (n-l+1)R-1$. In (95) $1(\cdot)$ is the indicator function, taking the value one if the argument is true, and zero if it is false. We get (96) by limiting ρ to the range $0\leq\rho\leq 1$ since the arguments of the minimization are both positive and upper-bounded by one. We use the i.i.d. property of the source, exchanging sums and products to get (97). The bound in (98) is true for all ρ in the range $0\leq\rho\leq 1$. Maximizing (98) over ρ gives $p_n(l)\leq\exp\{-(n-l+1)E_{pt,x}(R)\}$ where $E_{pt,x}(R)\}$ is defined in Theorem 2, in particular in (16).

B. Proof of Lemma 3

In this section we provide the proof of Lemma 3. We refer to (99)–(106) in the following discussion. The bound depends on whether $l\leq k$ or $l\geq k$. Consider the case $l\leq k$, In (99) we explicitly indicate the three conditions that a suffix pair $(\tilde{x}_l^n,\tilde{y}_k^n)$ must satisfy to result in a decoding error. In (100) we sum out over the common prefixes (x^{l-1},y^{l-1}), and use the fact that the random binning is done independently at each encoder, see Definition. 2. We get (101) by limiting ρ to the interval $0\leq\rho\leq 1$, as in (96). Getting (102) from (101) follows by a number of basic manipulations. In (102) we get the single letter expression by again using the memoryless property of the sources. In (103) we use the definitions of $E_{x|y}$ and E_{xy} from (7) and (8) of Theorem 6. Noting that the bound holds for all $\rho\in[0,1]$ optimizing over ρ results in (105). Finally, using the definition of (29) gives (106). The bound on $p_n(l,k)$ when $l>k$, is developed in an analogous fashion.

C. Equivalence of the Two Forms of the Error Exponent for Streaming Slepian-Wolf

In this section we prove the following lemma.

Lemma 5:

$$E_x(R_x,R_y,\gamma)=\sup_{\rho\in[0,1]}\left\{\gamma\,E_{x|y}(R_x,\rho)+(1-\gamma)E_{xy}(R_x,R_y,\rho)\right\} \tag{107}$$

$$=\inf_{q_{xy},o_{xy}}\Big\{\gamma\,D(q_{xy}||p_{xy})+(1-\gamma)D(o_{xy}||p_{xy}) + \max\{0,\gamma\,(R_x-H(q_{x|y}))+(1-\gamma)(R_x+R_y-H(o_{xy}))\}\Big\}, \tag{108}$$

where $E_{x|y}(\cdot)$ and $E_{xy}(\cdot)$ are defined in (7) and (8). For notational simplicity, we write q_{xy} and o_{xy} as two arbitrary joint distributions on $\mathcal{X}\times\mathcal{Y}$ (instead of $p_{\bar{x}\bar{y}}$ and $p_{\bar{\bar{x}}\bar{\bar{y}}}$). We retain p_{xy} to indicate the joint distribution of the source. The equivalence between the forms of $E_y(R_x,R_y,\gamma)$ in (29) and (30) can be proved using the same approach.

The proof of Lemma 5 is given in Section E. We start by giving some preliminary definitions in Section D. The proofs of a number of technical lemmas are deferred to Section F.

D. Preliminaries

We recall that the first form of the exponent specified in (107) resulted from the analysis of ML decoding. To help establish the lemma we define the function $E_{ml,x}(R_x,R_y,\gamma,\rho)$ as

$$\begin{aligned}E_{ml,x}(R_x,R_y,\gamma,\rho)&=\gamma\,E_{x|y}(R_x,\rho)+(1-\gamma)E_{xy}(R_x,R_y,\rho)\\&=\rho R^{(\gamma)}-\gamma\log\left(\sum_y\left(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)\\&\quad-(1-\gamma)(1+\rho)\log\left(\sum_y\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\right),\end{aligned}$$

and define

$$E_{ml,x}(R_x,R_y,\gamma)=\sup_{\rho\in[0,1]}E_{ml,x}(R_x,R_y,\gamma,\rho).$$

In addition, we use $E_{un,x}(R_x,R_y,\gamma)$ to denote the "universal" form (108) of the exponent, i.e.,

$$\begin{aligned}E_{un,x}(R_x,R_y,\gamma)&=\inf_{q_{xy},o_{xy}}\Big\{\gamma\,D(q_{xy}||p_{xy})+(1-\gamma)D(o_{xy}||p_{xy})\\&\quad+\max\left\{0,\gamma\,(R_x-H(q_{x|y}))+(1-\gamma)(R_x+R_y-H(o_{xy}))\right\}\Big\}\\&=\inf_{q_{xy},o_{xy}}\Big\{\gamma\,D(q_{xy}||p_{xy})+(1-\gamma)D(o_{xy}||p_{xy})\\&\quad+\max\{0,R^{(\gamma)}-\gamma\,H(q_{x|y})-(1-\gamma)H(o_{xy})\}\Big\}.\end{aligned}$$

To increase compactness we have defined

$$R^{(\gamma)}=\gamma\,R_x+(1-\gamma)(R_x+R_y).$$

We note that in the achievable rate region we have the relation

$$R^{(\gamma)}>\gamma\,H(p_{x|y})+(1-\gamma)H(p_{x,y}).$$

Finally, before starting the proof, we define a pair of distributions that we will need.

$$
\begin{aligned}
p_n(l,k) &= \sum_{x^n,y^n} p_{\mathbf{x},\mathbf{y}}(x^n,y^n)\Pr[\exists\ (\tilde{x}^n,\tilde{y}^n)\in\mathcal{B}_x(x^n)\times\mathcal{B}_y(y^n)\cap\mathcal{F}_n(l,k,x^n,y^n)\ \text{s.t.}\ p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)<p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^n,\tilde{y}_l^n)] \\
&\le \sum_{x^n,y^n}\min\Big[1,\sum_{\substack{(\tilde{x}^n,\tilde{y}^n)\in\mathcal{F}_n(l,k,x^n,y^n)\\ p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)<p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^n,\tilde{y}_l^n)}}\Pr[\tilde{x}^n\in\mathcal{B}_x(x^n),\tilde{y}^n\in\mathcal{B}_y(y^n)]\Big]p_{\mathbf{x},\mathbf{y}}(x^n,y^n) \qquad (99)\\
&\le \sum_{x_l^n,y_l^n}\min\Big[1,\sum_{\substack{(\tilde{x}_l^n,\tilde{y}_l^n)\ \text{s.t.}\ \tilde{y}^{k-1}=y^{k-1}\\ p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)<p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^n,\tilde{y}_l^n)}}\exp\{-(n-l+1)R_x-(n-k+1)R_y\}\Big]p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) \qquad (100)\\
&= \sum_{x_l^n,y_l^n}\min\Big[1,\sum_{\tilde{x}_l^n,\tilde{y}_k^n}\exp\{-(n-l+1)R_x-(n-k+1)R_y\} \\
&\qquad\qquad I[p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1})p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n)>p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)]\Big]p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) \\
&\le \sum_{x_l^n,y_l^n}\min\Big[1,\sum_{\tilde{x}_l^n,\tilde{y}_k^n}\exp\{-(n-l+1)R_x-(n-k+1)R_y\} \\
&\qquad\qquad \times\min\Big[1,\frac{p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1})p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n)}{p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)}\Big]\Big]p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) \\
&\le \sum_{x_l^n,y_l^n}\Big[\sum_{\tilde{x}_l^n,\tilde{y}_k^n}e^{-(n-l+1)R_x-(n-k+1)R_y}\Big[\frac{p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1})p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n)}{p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)}\Big]^{\frac{1}{1+\rho}}\Big]^{\rho}p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n) \qquad (101)\\
&= e^{-(n-l+1)\rho R_x-(n-k+1)\rho R_y}\sum_{x_l^n,y_l^n}\Big[\sum_{\tilde{x}_l^n,\tilde{y}_k^n}[p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1})p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n)]^{\frac{1}{1+\rho}}\Big]^{\rho}p_{\mathbf{x},\mathbf{y}}(x_l^n,y_l^n)^{\frac{1}{1+\rho}} \\
&= e^{-(n-l+1)\rho R_x-(n-k+1)\rho R_y}\sum_{y_l^{k-1}}\Big[\sum_{x_l^{k-1}}p_{\mathbf{x},\mathbf{y}}(x_l^{k-1},y_l^{k-1})^{\frac{1}{1+\rho}}\Big]\Big[\sum_{\tilde{x}_l^{k-1}}p_{\mathbf{x},\mathbf{y}}(\tilde{x}_l^{k-1},y_l^{k-1})^{\frac{1}{1+\rho}}\Big]^{\rho} \\
&\qquad\times\Big[\sum_{\tilde{x}_k^n,\tilde{y}_k^n}p_{\mathbf{x},\mathbf{y}}(\tilde{x}_k^n,\tilde{y}_k^n)^{\frac{1}{1+\rho}}\Big]^{\rho}\sum_{x_k^n,y_k^n}p_{\mathbf{x},\mathbf{y}}(x_k^n,y_k^n)^{\frac{1}{1+\rho}} \\
&= e^{-(n-l+1)\rho R_x-(n-k+1)\rho R_y}\Big[\sum_{y_l^{k-1}}\Big[\sum_{x_l^{k-1}}p_{\mathbf{x},\mathbf{y}}(x_l^{k-1},y_l^{k-1})^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big]\Big[\sum_{x_k^n,y_k^n}p_{\mathbf{x},\mathbf{y}}(x_k^n,y_k^n)^{\frac{1}{1+\rho}}\Big]^{1+\rho} \\
&= e^{-(n-l+1)\rho R_x-(n-k+1)\rho R_y}\Big[\sum_{y}\Big[\sum_{x}p_{x,y}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big]^{k-l}\Big[\sum_{x,y}p_{x,y}(x,y)^{\frac{1}{1+\rho}}\Big]^{(1+\rho)(n-k+1)} \qquad (102)\\
&= \exp\left\{-(k-l)\Big[\rho R_x-\log\Big[\sum_{y}\Big[\sum_{x}p_{x,y}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big]\Big]\right\} \\
&\qquad\times\exp\left\{-(n-k+1)\Big[\rho(R_x+R_y)-(1+\rho)\log\Big[\sum_{x,y}p_{x,y}(x,y)^{\frac{1}{1+\rho}}\Big]\Big]\right\} \\
&= \exp\big\{-(k-l)E_{x|y}(R_x,\rho)-(n-k+1)E_{xy}(R_x,R_y,\rho)\big\} \qquad (103)\\
&= \exp\left\{-(n-l+1)\Big[\frac{k-l}{n-l+1}E_{x|y}(R_x,\rho)+\frac{n-k+1}{n-l+1}E_{xy}(R_x,R_y,\rho)\Big]\right\} \qquad (104)\\
&\le \exp\left\{-(n-l+1)\sup_{\rho\in[0,1]}\Big[\frac{k-l}{n-l+1}E_{x|y}(R_x,\rho)+\frac{n-k+1}{n-l+1}E_{xy}(R_x,R_y,\rho)\Big]\right\} \qquad (105)\\
&= \exp\left\{-(n-l+1)E_x\left(R_x,R_y,\frac{k-l}{n-l+1}\right)\right\}. \qquad (106)
\end{aligned}
$$

Definition 5: Tilted distribution of p_{xy}: p_{xy}^{ρ}, for all $\rho\in[-1,\infty)$

$$p_{xy}^{\rho}(x,y)=\frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}. \qquad (109)$$

The entropy of the tilted distribution is written as $H(p_{xy}^{\rho})$. Obviously $p_{xy}^{0}=p_{xy}$.

Definition 6: The $x-y$ tilted distribution of p_{xy}, $\bar{p}_{xy}^{\rho}$, is defined for all $\rho\in[-1,+\infty)$ as

$$\bar{p}_{xy}^{\rho}(x,y)=\frac{\Big[\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}}{\sum_t\Big[\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}\Big]^{1+\rho}}\frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}}$$

$$= \frac{A(y,\rho)}{B(\rho)} \times \frac{C(x,y,\rho)}{D(y,\rho)}$$

where

$$A(y,\rho) = \left[\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}\right]^{1+\rho} = D(y,\rho)^{1+\rho},$$
$$B(\rho) = \sum_s \left[\sum_t p_{xy}(s,t)^{\frac{1}{1+\rho}}\right]^{1+\rho} = \sum_y A(y,\rho),$$
$$C(x,y,\rho) = p_{xy}(x,y)^{\frac{1}{1+\rho}},$$
$$D(y,\rho) = \sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}} = \sum_x C(x,y,\rho).$$

The marginal distribution for y is $\frac{A(y,\rho)}{B(\rho)}$. Obviously $\bar{p}^0_{xy} = p_{xy}$. Write the conditional distribution of x given y under distribution $\bar{p}^\rho_{xy}$ as $\bar{p}^\rho_{x|y}$, where $\bar{p}^\rho_{x|y}(x,y) = \frac{C(x,y,\rho)}{D(y,\rho)}$, and the conditional entropy of x given y under distribution $\bar{p}^\rho_{xy}$ as $H(\bar{p}^\rho_{x|y})$. Obviously $H(\bar{p}^0_{x|y}) = H(p_{x|y})$.

The conditional entropy of x given y for the $x-y$ tilted distribution is

$$H(\bar{p}^\rho_{x|y=y}) = -\sum_x \frac{C(x,y,\rho)}{D(y,\rho)} \log\left(\frac{C(x,y,\rho)}{D(y,\rho)}\right) \quad (110)$$

We introduce $A(y,\rho)$, $B(\rho)$, $C(x,y,\rho)$, $D(y,\rho)$ to simplify the notations. Some of their properties are shown in Lemma 9.

While tilted distributions are common optimal distributions in large deviation theory, it is useful to contemplate why we need to introduce these *two* tilted distributions. In the proof of Lemma 5 we show through a Lagrange multiplier argument that $\{p^\rho_{xy} : \rho \in [-1,+\infty)\}$ is the family of distributions that minimize the Kullback-Leibler distance to p_{xy} with fixed *entropy* and $\{\bar{p}^\rho_{xy} : \rho \in [-1,+\infty)\}$ is the family of distributions that minimize the Kullback−Leibler distance to p_{xy} with fixed *conditional entropy*. Using a Lagrange multiplier argument, we parametrize the universal error exponent $E_{un,x}(R_x, R_y, \gamma)$ in terms of ρ and show the equivalence of the universal and maximum likelihood error exponents.

E. Proof of Lemma 5

The proof of the lemma splits into two cases. Case 1 is when $\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) < R^{(\gamma)} < \gamma H(\bar{p}^1_{x|y}) + (1-\gamma)H(p^1_{xy})$. Case 2 is when $R^{(\gamma)} \geq \gamma H(\bar{p}^1_{x|y}) + (1-\gamma)H(p^1_{xy})$.

Proof:

Case 1: First, from Lemma F and Lemma 14:

$$\frac{\partial E_{ml,x}(R_x,R_y,\gamma,\rho)}{\partial\rho} = R^{(\gamma)} - \gamma H(\bar{p}^\rho_{x|y}) - (1-\gamma)H(p^\rho_{xy}). \quad (111)$$

Then, using Lemma 6 and Lemma 10, we have:

$$\frac{\partial^2 E_{ml,x}(R_x,R_y,\gamma,\rho)}{\partial\rho} \leq 0. \quad (112)$$

So ρ maximize $E_{ml,x}(R_x, R_y, \gamma, \rho)$, if and only if:

$$0 = \frac{\partial E_{ml,x}(R_x,R_y,\gamma,\rho)}{\partial\rho} = R^{(\gamma)} - \gamma H(\bar{p}^\rho_{x|y}) - (1-\gamma)H(p^\rho_{xy}).$$

Because $R^{(\gamma)}$ is in the interval $[\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}), \gamma H(\bar{p}^1_{x|y}) + (1-\gamma)H(p^1_{xy})]$ and the entropy functions monotonically-increase over ρ, we can find $\rho^* \in (0,1)$, s.t.

$$\gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy}) = R^{(\gamma)}.$$

Using Lemma 13 and Lemma F we get:

$$E_{ml,x}(R_x,R_y,\gamma) = \gamma D(\bar{p}^{\rho^*}_{xy}\|p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}\|p_{xy}). \quad (113)$$

Where $\gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy}) = R^{(\gamma)}$, ρ^* is generally unique because both $H(\bar{p}^\rho_{x|y})$ and $H(p^\rho_{xy})$ are strictly increasing with ρ.

Secondly,

$$\begin{aligned}
&E_{un,x}(R_x, R_y, \gamma)\\
&= \inf_{q_{xy},o_{xy}} \Big\{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy})\\
&\qquad + \max\left\{0, R^{(\gamma)} - \gamma H(q_{x|y}) - (1-\gamma)H(o_{xy})\right\}\Big\}\\
&= \inf_b \Big\{ \inf_{q_{xy},o_{xy}:\gamma H(q_{x|y})+(1-\gamma)H(o_{xy})=b} \left\{\gamma D(q_{xy}||p_{xy})\right.\\
&\qquad \left.+(1-\gamma)D(o_{xy}||p_{xy}) + \max(0, R^{(\gamma)} - b)\right\}\Big\}\\
&= \inf_{b\geq\gamma H(p_{x|y})+(1-\gamma)H(p_{xy})} \Big\{ \inf_{q_{xy},o_{xy}:\gamma H(q_{x|y})+(1-\gamma)H(o_{xy})=b}\\
&\qquad \left\{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) + \max(0, R^{(\gamma)} - b)\right\}\Big\}.
\end{aligned} \quad (114)$$

The last equality is true because, as we now show, $b < \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) < R^{(\gamma)}$ cannot be the optimizing choice of b. To see this note that the inner infimum in (114) is at least as large as $\max(0, R^{(\gamma)} - b)$, which can be lower bounded as:

$$\begin{aligned}
&\max(0, R^{(\gamma)} - b)\\
&\quad = \inf_{q_{xy},o_{xy}:H(q_{x|y})=H(p_{x|y}),H(o_{xy})=H(p_{xy})} \left\{\gamma D(q_{xy}||p_{xy})\right.\\
&\qquad \left.+(1-\gamma)D(o_{xy}||p_{xy}) + \max(0, R^{(\gamma)} - b)\right\}\\
&\quad \geq \inf_{q_{xy},o_{xy}:H(q_{x|y})=H(p_{x|y}),H(o_{xy})=H(p_{xy})} \left\{\gamma D(q_{xy}||p_{xy})\right.\\
&\qquad +(1-\gamma)D(o_{xy}||p_{xy})\\
&\qquad \left.+\max(0, R^{(\gamma)} - \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}))\right\}\\
&\quad \geq \inf_{q_{xy},o_{xy}:\gamma H(q_{x|y})+(1-\gamma)H(o_{xy})=\gamma H(p_{x|y})+(1-\gamma)H(p_{xy})}\\
&\qquad \Big\{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy})\\
&\qquad \left.+\max(0, R^{(\gamma)} - \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}))\right\},
\end{aligned}$$

which can be achieved in (114) through choosing $b = \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$. Hence, the minimizing b must be at least this large.

Now, fixing $b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$, the inner infimum in (114) is an optimization problem on q_{xy}, o_{xy} with equality constraints $\sum_x \sum_y q_{xy}(x,y) = 1$, $\sum_x \sum_y o_{xy}(x,y) = 1$ and $\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b$ and the obvious inequality constraints $0 \leq q_{xy}(x,y) \leq 1$, $0 \leq o_{xy}(x,y) \leq 1, \forall x, y$. In the following formulation of the

optimization problem, we relax one equality constraint to an inequality constraint $\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) \geq b$ to make the optimization problem *convex*. It turns out later that the optimal solution to the relaxed problem is also the optimal solution to the original problem because $b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$. The resulting optimization problem is:

$$\begin{aligned}
&\inf_{q_{xy}, o_{xy}} \quad \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy})\} \\
&\text{s.t.} \quad \sum_x \sum_y q_{xy}(x,y) = 1 \\
&\qquad \sum_x \sum_y o_{xy}(x,y) = 1 \\
&b - \gamma H(q_{x|y}) - (1-\gamma)H(o_{xy}) \leq 0 \\
&0 \leq q_{xy}(x,y) \leq 1, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \\
&0 \leq o_{xy}(x,y) \leq 1, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}
\end{aligned} \tag{115}$$

The above optimization problem is *convex* because the objective function and the inequality constraint functions are convex and the equality constraint functions are affine [1]. The Lagrange multiplier function for this convex optimization problem is:

$$\begin{aligned}
&L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4) \\
&\quad = \gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) \\
&\qquad + \mu_1 \Big(\sum_x \sum_y q_{xy}(x,y) - 1\Big) + \mu_2 \Big(\sum_x \sum_y o_{xy}(x,y) - 1\Big) \\
&\qquad + \rho(b - \gamma H(q_{x|y}) - (1-\gamma)H(o_{xy})) \\
&\qquad + \sum_x \sum_y \Big\{\nu_1(x,y)(-q_{xy}(x,y)) \\
&\qquad + \nu_2(x,y)(1 - q_{xy}(x,y)) + \nu_3(x,y)(-o_{xy}(x,y)) \\
&\qquad + \nu_4(x,y)(1 - o_{xy}(x,y))\Big\},
\end{aligned}$$

where ρ, μ_1, μ_2 are real numbers and $\nu_i \in R^{|\mathcal{X}||\mathcal{Y}|}$, $i = 1, 2, 3, 4$.

According to the KKT conditions for convex optimization [1], q_{xy}, o_{xy} minimize the convex optimization problem in (115) if and only if the following conditions are simultaneously satisfied for some q_{xy}, o_{xy}, μ_1, μ_2, ν_1, ν_2, ν_3, ν_4 and ρ:

$$\begin{aligned}
0 &= \frac{\partial L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4)}{\partial q_{xy}(x,y)} \\
&= \gamma \Big[-\log(p_{xy}(x,y)) + (1+\rho)(1 + \log(q_{xy}(x,y))) \\
&\qquad + \rho \log\Big(\sum_s q_{xy}(s,y)\Big)\Big] + \mu_1 - \nu_1(x,y) - \nu_2(x,y) \\
0 &= \frac{\partial L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4)}{\partial o_{xy}(x,y)} \\
&= (1-\gamma)\Big[-\log(p_{xy}(x,y)) + (1+\rho)(1 + \log(o_{xy}(x,y)))\Big] \\
&\qquad + \mu_2 - \nu_3(x,y) - \nu_4(x,y)
\end{aligned} \tag{116}$$

For all x, y and

$$\begin{aligned}
\rho\Big(\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) - b\Big) &= 0 \\
\rho &\geq 0 \\
\nu_1(x,y)(-q_{xy}(x,y)) &= 0, \ \forall x, y \\
\nu_2(x,y)(1 - q_{xy}(x,y)) &= 0, \ \forall x, y \\
\nu_3(x,y)(-o_{xy}(x,y)) &= 0, \ \forall x, y \\
\nu_4(x,y)(1 - o_{xy}(x,y)) &= 0, \ \forall x, y \\
\nu_i(x,y) &\geq 0, \ \forall x, y, 1 \leq i \leq 4
\end{aligned} \tag{117}$$

Solving the above standard Lagrange multiplier equations (116) and (117), we have:

$$\begin{aligned}
q_{xy}(x,y) &= \frac{[\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho_b}}]^{1+\rho_b}}{\sum_t [\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho_b}}]^{1+\rho_b}} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho_b}}}{\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho_b}}} \\
&= \bar{p}^{\rho_b}_{xy}(x,y) \\
o_{xy}(x,y) &= \frac{p_{xy}(x,y)^{\frac{1}{1+\rho_b}}}{\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho_b}}} \\
&= p^{\rho_b}_{xy}(x,y) \\
\nu_i(x,y) &= 0 \quad \forall x, y, i = 1, 2, 3, 4 \\
\rho &= \rho_b
\end{aligned} \tag{118}$$

Where ρ_b satisfies the condition

$$\begin{aligned}
\gamma H(\bar{p}^{\rho_b}_{x|y}) + (1-\gamma)H(p^{\rho_b}_{xy}) = b \geq \gamma H(p_{x|y}) \\
+ (1-\gamma)H(p_{xy})
\end{aligned} \tag{119}$$

and thus $\rho_b \geq 0$ because both $H(\bar{p}^{\rho}_{x|y})$ and $H(p^{\rho}_{xy})$ are monotonically increasing with ρ as shown in Lemma 6 and Lemma 10.

Notice that all the KKT conditions are simultaneously satisfied with the inequality constraint $\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) \geq b$ being met with equality. Thus, the relaxed optimization problem has the same optimal solution as the original problem as promised. The optimal q_{xy} and o_{xy} are the $x-y$ tilted distribution $\bar{p}^{\rho_b}_{xy}$ and standard tilted distribution $p^{\rho_b}_{xy}$ of p_{xy} with the same parameter $\rho_b \geq 0$. chosen s.t.

$$\gamma H(\bar{p}^{\rho_b}_{x|y}) + (1-\gamma)H(p^{\rho_b}_{xy}) = b \tag{120}$$

Now, consider the expansion of $E_{un,x}(R_x, R_y, \gamma)$ provided in (121). Later in the appendix we show that $H(p^{\rho}_{xy})$, $H(\bar{p}^{\rho}_{x|y})$, $D(\bar{p}^{\rho}_{xy}||p_{xy})$ and $D(p^{\rho}_{xy}||p_{xy})$ are all strictly increasing with $\rho > 0$, shown respectively in Lemma 10, Lemma 11, Lemma 6 and Lemma 7. Now, consider each term of (121). The second term simplifies to

$$\begin{aligned}
&\inf_{\rho \geq 0 : R^{(\gamma)} \leq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) \\
&\qquad + (1-\gamma)D(p^{\rho}_{xy}||p_{xy})\} \\
&\quad = \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy})
\end{aligned} \tag{122}$$

where $R^{(\gamma)} = \gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy})$.

$$
\begin{aligned}
&E_{un,x}(R_x, R_y, \gamma) \\
&\quad= \inf_{b \geq \gamma H(p_{x|y})+(1-\gamma)H(p_{xy})} \left\{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y})+(1-\gamma)H(o_{xy})=b} \left\{ \gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + \max(0, R^{(\gamma)} - b) \right\} \right\} \\
&\quad= \inf_{b \geq \gamma H(p_{x|y})+(1-\gamma)H(p_{xy})} \left\{ \gamma D(\bar{p}_{xy}^{\rho_b}||p_{xy}) + (1-\gamma) D(p_{xy}^{\rho_b}||p_{xy}) + \max(0, R^{(\gamma)} - b) \right\} \\
&\quad= \min \left[\inf_{\rho \geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{\rho})+(1-\gamma)H(p_{xy}^{\rho})} \left\{ \gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) + (1-\gamma) D(p_{xy}^{\rho}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma) H(p_{xy}^{\rho}) \right\}, \right. \\
&\qquad \left. \inf_{\rho \geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}_{x|y}^{\rho})+(1-\gamma)H(p_{xy}^{\rho})} \left\{ \gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) + (1-\gamma) D(p_{xy}^{\rho}||p_{xy}) \right\} \right] \qquad (121)
\end{aligned}
$$

Applying the results of Lemma 12 and Lemma 8 to the first term of (121) we get:

$$
\begin{aligned}
&\inf_{\rho \geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{\rho})+(1-\gamma)H(p_{xy}^{\rho})} \left\{ \gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) \right. \\
&\left. +(1-\gamma) D(p_{xy}^{\rho}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma) H(p_{xy}^{\rho}) \right\} \\
&= \left[\gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) + (1-\gamma) D(p_{xy}^{\rho}||p_{xy}) \right. \\
&\quad \left. + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma) H(p_{xy}^{\rho}) \right] \Big|_{\rho=\rho^*} \\
&= \gamma D(\bar{p}_{xy}^{\rho^*}||p_{xy}) + (1-\gamma) D(p_{xy}^{\rho^*}||p_{xy}). \qquad (123)
\end{aligned}
$$

This is true because for $\rho : R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma) H(p_{xy}^{\rho})$, we know $\rho \leq 1$ because of the range of $R^{(\gamma)}$: $R^{(\gamma)} < \gamma H(\bar{p}_{x|y}^{1}) + (1-\gamma) H(p_{xy}^{1})$. Substituting (122) and (123) into (121), we get

$$
\begin{aligned}
E_{un,x}(R_x, R_y, \gamma) &= \gamma D(\bar{p}_{xy}^{\rho^*}||p_{xy}) + (1-\gamma) D(p_{xy}^{\rho^*}||p_{xy}) \\
\text{where} \quad R^{(\gamma)} &= \gamma H(\bar{p}_{x|y}^{\rho^*}) + (1-\gamma) H(p_{xy}^{\rho^*}). \qquad (124)
\end{aligned}
$$

So for $\gamma H(p_{x|y}) + (1-\gamma) H(p_{xy}) \leq R^{(\gamma)} \leq \gamma H(\bar{p}_{x|y}^{1}) + (1-\gamma) H(p_{xy}^{1})$, from (113) we have the desired property:

$$
E_{ml,x}(R_x, R_y, \gamma) = E_{un,x}(R_x, R_y, \gamma). \qquad (125)
$$

Case 2: Recall that this is the case where $R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{1}) + (1-\gamma) H(p_{xy}^{1})$. In this case, for all $0 \leq \rho \leq 1$

$$
\begin{aligned}
\frac{\partial E_{ml,x}(R_x, R_y, \gamma, \rho)}{\partial \rho} &= R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma) H(p_{xy}^{\rho}) \\
&\geq R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{1}) - (1-\gamma) H(p_{xy}^{1}) \geq 0.
\end{aligned}
$$

So ρ takes value 1 to maximize the error exponent $E_{ml,x}(R_x, R_y, \gamma, \rho)$, thus

$$
\begin{aligned}
E_{ml,x}(R_x, R_y, \gamma) &= R^{(\gamma)} - \gamma \log \Big(\sum_y \big(\sum_x p_{xy}(x,y)^{\frac{1}{2}} \big)^2 \Big) \\
&\quad -2(1-\gamma) \log \big(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{2}} \big).
\end{aligned}
$$

Using the same convex optimization techniques as case 1, we notice the fact that $\rho^* \geq 1$ for $R^{(\gamma)} = \gamma H(\bar{p}_{x|y}^{\rho^*}) + (1-\gamma) H(p_{xy}^{\rho^*})$. Then applying Lemma 12 and Lemma 8, we have:

$$
\begin{aligned}
&\inf_{\rho \geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{\rho})+(1-\gamma)H(p_{xy}^{\rho})} \left\{ \gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) \right. \\
&\left. +(1-\gamma) D(p_{xy}^{\rho}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma) H(p_{xy_\rho}) \right\}, \\
&= \gamma D(\bar{p}_{xy}^{1}||p_{xy}) + (1-\gamma) D(p_{xy}^{1}||p_{xy}) \\
&\quad + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{1}) - (1-\gamma) H(p_{xy}^{1}).
\end{aligned}
$$

And

$$
\begin{aligned}
&\inf_{\rho \geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}_{x|y}^{\rho})+(1-\gamma)H(p_{xy}^{\rho})} \{ \gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) \\
&\qquad\qquad +(1-\gamma) D(p_{xy}^{\rho}||p_{xy}) \} \\
&\quad = \gamma D(\bar{p}_{xy}^{\rho^*}||p_{xy}) + (1-\gamma) D(p_{xy}^{\rho^*}||p_{xy}) \\
&\quad = \gamma D(\bar{p}_{xy}^{\rho^*}||p_{xy}) + (1-\gamma) D(p_{xy}^{\rho^*}||p_{xy}) \qquad (126) \\
&\qquad + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho^*}) - (1-\gamma) H(p_{xy}^{\rho^*}) \\
&\quad \leq \gamma D(\bar{p}_{xy}^{1}||p_{xy}) + (1-\gamma) D(p_{xy}^{1}||p_{xy}) \qquad (127) \\
&\qquad + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{1}) - (1-\gamma) H(p_{xy}^{1})
\end{aligned}
$$

Finally, in (128)–(129) we complete the derivation. where the equality in (128) is true by setting $\rho = 1$ in Lemma 13 and Lemma F.

Thus, as for case 1, for this case we again find that $E_{ml,x}(R_x, R_y, \gamma) = E_{un,x}(R_x, R_y, \gamma)$, finishing the proof.

F. Proofs of Lemmas Used to Prove Lemma 5

Lemma 6: $\frac{\partial H(p_{xy}^{\rho})}{\partial \rho} \geq 0$

Proof: From the definition of the tilted distribution we have the following observation:

$$
\begin{aligned}
&\log(p_{xy}^{\rho}(x_1, y_1)) - \log(p_{xy}^{\rho}(x_2, y_2)) \\
&\qquad = \log(p_{xy}(x_1, y_1)^{\frac{1}{1+\rho}}) - \log(p_{xy}(x_2, y_2)^{\frac{1}{1+\rho}}).
\end{aligned}
$$

Using the above equality, we first derive the derivative of the tilted distribution, for all x, y

$$
\begin{aligned}
&\frac{\partial p_{xy}^{\rho}(x,y)}{\partial \rho} \\
&= \frac{-1}{(1+\rho)^2} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}} \log(p_{xy}(x,y)) (\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})}{(\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})^2} \\
&\quad - \frac{-1}{(1+\rho)^2} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}} (\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}} \log(p_{xy}(s,t)))}{(\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})^2}
\end{aligned}
$$

$$
\begin{aligned}
&E_{un,x}(R_x, R_y, \gamma) \\
&= \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma) H(p_{xy})} \left\{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma) H(o_{xy}) = b} \left\{ \gamma D(q_{xy} || p_{xy}) + (1-\gamma) D(o_{xy} || p_{xy}) + \max(0, R^{(\gamma)} - b) \right\} \right\} \quad (128) \\
&= \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma) H(p_{xy})} \left\{ \gamma D(\bar{p}_{xy}^{\rho_b} || p_{xy}) + (1-\gamma) D(p_{xy}^{\rho_b} || p_{xy}) + \max(0, R^{(\gamma)} - b) \right\} \\
&= \min \Big[\inf_{\rho \geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma) H(p_{xy}^{\rho})} \left\{ \gamma D(\bar{p}_{xy}^{\rho} || p_{xy}) + (1-\gamma) D(p_{xy}^{\rho} || p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma) H(p_{xy}^{\rho}) \right\}, \\
&\qquad \inf_{\rho \geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma) H(p_{xy}^{\rho})} \left\{ \gamma D(\bar{p}_{xy}^{\rho} || p_{xy}) + (1-\gamma) D(p_{xy}^{\rho} || p_{xy}) \right\} \Big] \\
&= \gamma D(\bar{p}_{xy}^{1} || p_{xy}) + (1-\gamma) D(p_{xy}^{1} || p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{1}) - (1-\gamma) H(p_{xy}^{1}) \\
&= R^{(\gamma)} - \gamma \log \Big(\sum_y \big(\sum_x p_{xy}(x,y)^{\frac{1}{2}} \big)^2 \Big) - 2(1-\gamma) \log \Big(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{2}} \Big), \quad (129)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{-1}{1+\rho} p_{xy}^{\rho}(x,y) \Big[\log(p_{xy}(x,y)^{\frac{1}{1+\rho}}) \\
&\qquad - \sum_t \sum_s p_{xy}^{\rho}(s,t) \log(p_{xy}(s,t)^{\frac{1}{1+\rho}}) \Big] \\
&= \frac{-1}{1+\rho} p_{xy}^{\rho}(x,y) \Big[\log(p_{xy}^{\rho}(x,y)) \\
&\qquad - \sum_t \sum_s p_{xy}^{\rho}(s,t) \log(p_{xy}^{\rho}(s,t)) \Big] \\
&= -\frac{p_{xy}^{\rho}(x,y)}{1+\rho} [\log(p_{xy}^{\rho}(x,y)) + H(p_{xy}^{\rho})]
\end{aligned}
$$

Then:

$$
\begin{aligned}
&\frac{\partial H(p_{xy}^{\rho})}{\partial \rho} \quad (130) \\
&= -\frac{\partial \sum_{x,y} p_{xy}^{\rho}(x,y) \log(p_{xy}^{\rho}(x,y))}{\partial \rho} \\
&= -\sum_{x,y} (1 + \log(p_{xy}^{\rho}(x,y))) \frac{\partial p_{xy}^{\rho}(x,y)}{\partial \rho} \\
&= \sum_{x,y} (1 + \log(p_{xy}^{\rho}(x,y))) \frac{p_{xy}^{\rho}(x,y)}{1+\rho} \\
&\qquad \times (\log(p_{xy}^{\rho}(x,y)) + H(p_{xy}^{\rho})) \\
&= \frac{1}{1+\rho} \sum_{x,y} p_{xy}^{\rho}(x,y) \log(p_{xy}^{\rho}(x,y)) \\
&\qquad \times (\log(p_{xy}^{\rho}(x,y)) + H(p_{xy}^{\rho})) \\
&= \frac{1}{1+\rho} \Big[\sum_{x,y} p_{xy}^{\rho}(x,y) (\log(p_{xy}^{\rho}(x,y)))^2 - H(p_{xy}^{\rho})^2 \Big] \\
&= \frac{1}{1+\rho} \Big[\sum_{x,y} p_{xy}^{\rho}(x,y) (\log(p_{xy}^{\rho}(x,y)))^2 \sum_{x,y} p_{xy}^{\rho}(x,y) \\
&\qquad - H(p_{xy}^{\rho})^2 \Big] \\
&\geq \frac{1}{1+\rho} \Big[\Big(\sum_{x,y} p_{xy}^{\rho}(x,y) \log(p_{xy}^{\rho}(x,y)) \Big)^2 - H(p_{xy}^{\rho})^2 \Big] \quad (131) \\
&= 0
\end{aligned}
$$

where (131) is true by the Cauchy-Schwartz inequality.

Lemma 7: $\frac{\partial D(p_{xy}^{\rho} \| P)}{\partial \rho} = \rho \frac{\partial H(p_{xy}^{\rho})}{\partial \rho}$

Proof: As shown in Lemma 13 and Lemma F respectively:

$$
D(p_{xy}^{\rho} \| p_{xy}) = \rho H(p_{xy}^{\rho}) - (1+\rho) \log(\sum_{x,y} p_{xy}(x,y)^{\frac{1}{1+\rho}})
$$

$$
H(p_{xy}^{\rho}) = \frac{\partial (1+\rho) \log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial \rho}
$$

We have:

$$
\begin{aligned}
&\frac{\partial D(p_{xy}^{\rho} \| p_{xy})}{\partial \rho} \\
&\qquad = H(p_{xy}^{\rho}) + \rho \frac{\partial H(p_{xy}^{\rho})}{\partial \rho} \\
&\qquad\quad - \frac{\partial (1+\rho) \log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial \rho} \\
&\qquad = H(p_{xy}^{\rho}) + \rho \frac{\partial H(p_{xy}^{\rho})}{\partial \rho} - H(p_{xy}^{\rho}) \\
&\qquad = \rho \frac{\partial H(p_{xy}^{\rho})}{\partial \rho} \quad (132)
\end{aligned}
$$

■

Lemma 8: $sign \frac{\partial [D(p_{xy}^{\rho} \| p_{xy}) - H(p_{xy}^{\rho})]}{\partial \rho} = sign(\rho - 1)$.

Proof: Combining the results of the previous two lemmas, we have:

$$
\begin{aligned}
\frac{\partial D(p_{xy}^{\rho} \| p_{xy}) - H(p_{xy}^{\rho})}{\partial \rho} &= (\rho - 1) \frac{\partial H(p_{xy}^{\rho})}{\partial \rho} \\
&= sign(\rho - 1)
\end{aligned}
$$

■

Lemma 9: Properties of $\frac{\partial A(y,\rho)}{\partial \rho}$, $\frac{\partial B(\rho)}{\partial \rho}$, $\frac{\partial C(x,y,\rho)}{\partial \rho}$, $\frac{\partial D(y,\rho)}{\partial \rho}$ and $\frac{\partial H(\bar{p}_{x|y=y}^{\rho})}{\partial \rho}$

First,

$$
\begin{aligned}
\frac{\partial C(x,y,\rho)}{\partial \rho} &= \frac{\partial p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\partial \rho} \\
&= -\frac{1}{1+\rho} p_{xy}(x,y)^{\frac{1}{1+\rho}} \log(p_{xy}(x,y)^{\frac{1}{1+\rho}}) \\
&= -\frac{C(x,y,\rho)}{1+\rho} \log(C(x,y,\rho)).
\end{aligned}
$$

$$\begin{aligned}\frac{\partial D(y,\rho)}{\partial\rho} &= \frac{\partial \sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}}{\partial\rho}\\ &= -\frac{1}{1+\rho}\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}\log(p_{xy}(s,y)^{\frac{1}{1+\rho}})\\ &= -\frac{\sum_x C(x,y,\rho)\log(C(x,y,\rho))}{1+\rho}.\end{aligned}$$

For a differentiable function $f(\rho)$,

$$\frac{\partial f(\rho)^{1+\rho}}{\partial\rho} = f(\rho)^{1+\rho}\log(f(\rho)) + (1+\rho)f(\rho)^{\rho}\frac{\partial f(\rho)}{\partial\rho}. \tag{133}$$

So

$$\begin{aligned}\frac{\partial A(y,\rho)}{\partial\rho} &= \frac{\partial D(y,\rho)^{1+\rho}}{\partial\rho}\\ &= D(y,\rho)^{1+\rho}\log(D(y,\rho))\\ &\quad +(1+\rho)D(y,\rho)^{\rho}\frac{\partial D(y,\rho)}{\partial\rho}\\ &= D(y,\rho)^{1+\rho}(\log(D(y,\rho))\\ &\quad -\sum_x \frac{C(x,y,\rho)}{D(y,\rho)}\log(C(x,y,\rho)))\\ &= D(y,\rho)^{1+\rho}(-\sum_x \frac{C(x,y,\rho)}{D(y,\rho)}\log(\frac{C(x,y,\rho)}{D(y,\rho)}))\\ &= A(y,\rho)H(\bar{p}^{\rho}_{x|y=y})\\ \frac{\partial B(\rho)}{\partial\rho} &= \sum_y \frac{\partial A(y,\rho)}{\partial\rho} = \sum_y A(y,\rho)H(\bar{p}^{\rho}_{x|y=y})\\ &= B(\rho)\sum_y \frac{A(y,\rho)}{B(\rho)}H(\bar{p}^{\rho}_{x|y=y}) = B(\rho)H(\bar{p}^{\rho}_{x|y})\end{aligned}$$

And last:

$$\begin{aligned}&\frac{\partial H(\bar{p}^{\rho}_{x|y=y})}{\partial\rho}\\ &=-\sum_x\left[\frac{\frac{\partial C(x,y,\rho)}{\partial\rho}}{D(y,\rho)} - \frac{C(x,y,\rho)\frac{\partial D(y,\rho)}{\partial\rho}}{D(y,\rho)^2}\right]\left[1+\log\frac{C(x,y,\rho)}{D(y,\rho)}\right]\\ &= -\sum_x\left[\frac{-\frac{C(x,y,\rho)}{1+\rho}\log(C(x,y,\rho))}{D(y,\rho)}\right.\\ &\quad\left.+\frac{C(x,y,\rho)\frac{\sum_s C(s,y,\rho)\log(C(s,y,\rho))}{1+\rho}}{D(y,\rho)^2}\right]\left[1+\log\frac{C(x,y,\rho)}{D(y,\rho)}\right]\\ &= \frac{1}{1+\rho}\sum_x\Big[\bar{p}^{\rho}_{x|y}(x,y)\log(C(x,y,\rho))\\ &\quad -\bar{p}^{\rho}_{x|y}(x,y)\sum_s\bar{p}^{\rho}_{x|y}(s,y)\log(C(s,y,\rho))\Big]\Big[1+\log(\bar{p}^{\rho}_{x|y}(x,y))\Big]\\ &= \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{x|y}(x,y)\Big[\log(\bar{p}^{\rho}_{x|y}(x,y))\\ &\quad -\sum_s \bar{p}^{\rho}_{x|y}(s,y)\log(\bar{p}^{\rho}_{x|y}(s,y))\Big]\Big[1+\log(\bar{p}^{\rho}_{x|y}(x,y))\Big]\\ &= \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{x|y}(x,y)\log(\bar{p}^{\rho}_{x|y}(x,y))\Big[\log(\bar{p}^{\rho}_{x|y}(x,y))\\ &\quad -\sum_s \bar{p}^{\rho}_{x|y}(s,y)\log(\bar{p}^{\rho}_{x|y}(s,y))\Big]\\ &= \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{x|y}(x,y)\log(\bar{p}^{\rho}_{x|y}(x,y))\log(\bar{p}^{\rho}_{x|y}(x,y))\\ &\quad -\frac{1}{1+\rho}\left[\sum_x \bar{p}^{\rho}_{x|y}(x,y)\log(\bar{p}^{\rho}_{x|y}(x,y))\right]^2\\ &\geq 0\end{aligned}$$

The inequality is true by the Cauchy-Schwartz inequality and by noticing that $\sum_x \bar{p}^{\rho}_{x|y}(x,y) = 1$. ■

These properties will again be used in the proofs in the following lemmas.

Lemma 10: $\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial\rho} \geq 0$

Proof:

$$\begin{aligned}\frac{\partial\frac{A(y,\rho)}{B(\rho)}}{\partial\rho} &= \frac{1}{B(\rho)^2}\left[\frac{\partial A(y,\rho)}{\partial\rho}B(\rho) - \frac{\partial B(\rho)}{\partial\rho}A(y,\rho)\right]\\ &= \frac{1}{B(\rho)^2}\Big[A(y,\rho)H\left(\bar{p}^{\rho}_{x|y=y}\right)B(\rho)\\ &\quad - H\left(\bar{p}^{\rho}_{x|y}\right)B(\rho)A(y,\rho)\Big]\\ &= \frac{A(y,\rho)}{B(\rho)}\left[H(\bar{p}^{\rho}_{x|y=y}) - H(\bar{p}^{\rho}_{x|y})\right]\end{aligned}$$

Now,

$$\begin{aligned}\frac{\partial H\left(\bar{p}^{\rho}_{x|y}\right)}{\partial\rho} &= \frac{\partial}{\partial\rho}\sum_y\frac{A(y,\rho)}{B(\rho)}\sum_x\frac{C(x,y,\rho)}{D(y,\rho)}\left[-\log\frac{C(x,y,\rho)}{D(y,\rho)}\right]\\ &= \frac{\partial}{\partial\rho}\sum_y\frac{A(y,\rho)}{B(\rho)}H\left(\bar{p}^{\rho}_{x|y=y}\right)\\ &= \sum_y\frac{A(y,\rho)}{B(\rho)}\frac{\partial H\left(\bar{p}^{\rho}_{x|y=y}\right)}{\partial\rho}\\ &\quad +\sum_y\frac{\partial\frac{A(y,\rho)}{B(\rho)}}{\partial\rho}H\left(\bar{p}^{\rho}_{x|y=y}\right)\\ &\geq \sum_y\frac{\partial\frac{A(y,\rho)}{B(\rho)}}{\partial\rho}H\left(\bar{p}^{\rho}_{x|y=y}\right)\\ &= \sum_y\frac{A(y,\rho)}{B(\rho)}\left[H\left(\bar{p}^{\rho}_{x|y=y}\right) - H\left(\bar{p}^{\rho}_{x|y}\right)\right]\\ &\quad \times H\left(\bar{p}^{\rho}_{x|y=y}\right)\\ &= \sum_y\frac{A(y,\rho)}{B(\rho)}H\left(\bar{p}^{\rho}_{x|y=y}\right)^2 - H\left(\bar{p}^{\rho}_{x|y}\right)^2\\ &= \left[\sum_y\frac{A(y,\rho)}{B(\rho)}H\left(\bar{p}^{\rho}_{x|y=y}\right)^2\right]\left[\sum_y\frac{A(y,\rho)}{B(\rho)}\right]\\ &\quad -H\left(\bar{p}^{\rho}_{x|y}\right)^2\\ &\geq \left[\sum_y\frac{A(y,\rho)}{B(\rho)}H\left(\bar{p}^{\rho}_{x|y=y}\right)\right]^2 - H\left(\bar{p}^{\rho}_{x|y}\right)^2\\ &= 0\end{aligned}$$

where the last inequality is again true by Cauchy-Schwartz.

Lemma 11: $\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy})}{\partial\rho}=\rho\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial\rho}$

Proof: As shown in Lemma F and Lemma 14 respectively:

$$D(\bar{p}^{\rho}_{xy}\|p_{xy})=\rho H(\bar{p}^{\rho}_{x|y})-\log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})$$

$$H(\bar{p}^{\rho}_{x|y})=\frac{\partial\log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial\rho}$$

We have:

$$\begin{aligned}&\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy})}{\partial\rho}\\&\quad=H(\bar{p}^{\rho}_{x|y})+\rho\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial\rho}\\&\qquad-\frac{\partial\log\left(\sum_y\left(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)}{\partial\rho}\\&\quad=H(\bar{p}^{\rho}_{x|y})+\rho\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial\rho}-H(\bar{p}^{\rho}_{x|y})\\&\quad=\rho\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial\rho}\end{aligned}$$

Lemma 12: $sign\frac{\partial[D(\bar{p}^{\rho}_{xy}\|p_{xy})-H(\bar{p}^{\rho}_{x|y})]}{\partial\rho}=sign(\rho-1)$.

Proof: Using the previous lemma, we get:

$$\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy})-H(\bar{p}^{\rho}_{x|y})}{\partial\rho}=(\rho-1)\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial\rho}$$

Then by Lemma 10, we get the conclusion. ■

Lemma 13:

$$\rho H(p^{\rho}_{xy})-(1+\rho)\log\left(\sum_y\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\right)=D(p^{\rho}_{xy}\|p_{xy}).\tag{134}$$

Proof: By noticing that

$$\begin{aligned}&\log(p_{xy}(x,y))\\&\quad=(1+\rho)\left[\log(p^{\rho}_{xy}(x,y))+\log\left(\sum_{s,t}p_{xy}(s,t)^{\frac{1}{1+\rho}}\right)\right]\end{aligned}$$

we have:

$$\begin{aligned}&D(p^{\rho}_{xy}\|p_{xy})+H(p^{\rho}_{xy})=-\sum_{x,y}p^{\rho}_{xy}(x,y)\log(p_{xy}(x,y))\\&=-\sum_{x,y}p^{\rho}_{xy}(x,y)(1+\rho)\\&\quad\times\left[\log(p^{\rho}_{xy}(x,y))+\log\left(\sum_{s,t}p_{xy}(s,t)^{\frac{1}{1+\rho}}\right)\right]\\&=(1+\rho)H(p^{\rho}_{xy})-(1+\rho)\\&\quad\times\sum_{x,y}p^{\rho}_{xy}(x,y)\log\left(\sum_{s,t}p_{xy}(s,t)^{\frac{1}{1+\rho}}\right).\end{aligned}$$

And so,

$$D(p^{\rho}_{xy}\|p_{xy})=\rho H(p^{\rho}_{xy})-(1+\rho)\log\left(\sum_{s,t}p_{xy}(s,t)^{\frac{1}{1+\rho}}\right).$$

■

Lemma 14:

$$\rho H(\bar{p}^{\rho}_{x|y})-\log\left(\sum_y\left(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)=D(\bar{p}^{\rho}_{xy}\|p_{xy}).\tag{135}$$

$$\begin{aligned}D(\bar{p}^{\rho}_{xy}\|p_{xy})&=\sum_y\sum_x\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}\\&\quad\times\log\left(\frac{\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}}{p_{xy}(x,y)}\right)\\&=\sum_y\sum_x\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}\Big[\log\left(\frac{A(y,\rho)}{B(\rho)}\right)\\&\quad+\log\left(\frac{C(x,y,\rho)}{D(y,\rho)}\right)-\log\left(p_{xy}(x,y)\right)\Big]\\&=-\log(B(\rho))-H(\bar{p}^{\rho}_{x|y})\\&\quad+\sum_y\sum_x\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}\\&\quad\times\Big[\log(D(y,\rho)^{1+\rho})-\log(C(x,y,\rho)^{1+\rho})\Big]\\&=-\log(B(\rho))-H(\bar{p}^{\rho}_{x|y})+(1+\rho)H(\bar{p}^{\rho}_{x|y})\\&=-\log\left(\sum_y\left(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)\\&\quad+\rho H(\bar{p}^{\rho}_{x|y}).\end{aligned}$$

■

Lemma 15:

$$H(p^{\rho}_{xy})=\frac{\partial(1+\rho)\log(\sum_y\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial\rho}.\tag{136}$$

Proof:

$$\begin{aligned}&\frac{\partial(1+\rho)\log(\sum_y\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial\rho}\\&=\log\left(\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}\right)\\&\quad-\sum_y\sum_x\frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}\log\left(p_{xy}(x,y)^{\frac{1}{1+\rho}}\right)\\&=-\sum_y\sum_x\frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}\log\left(\frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}\right)\\&=H(p^{\rho}_{xy}).\end{aligned}$$

■

Lemma 14:

$$H(\bar{p}^{\rho}_{x|y})=\frac{\partial\log\left(\sum_y\left(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)}{\partial\rho}\tag{137}$$

Proof: Notice that $B(\rho)=\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}$, and $\frac{\partial B(\rho)}{\partial\rho}=B(\rho)H(\bar{p}^{\rho}_{x|y})$ as shown in Lemma 9. It is clear

that:

$$\begin{aligned}\frac{\partial \log\left(\sum_y \left(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right)}{\partial \rho} &= \frac{\partial \log(B(\rho))}{\partial \rho} \\ &= \frac{1}{B(\rho)}\frac{\partial B(\rho)}{\partial \rho} \\ &= H(\bar{p}^{\rho}_{x|y}).\end{aligned}$$

■

ACKNOWLEDGMENT

The authors wish to acknowledge a discussion with Professor Zixiang Xiong during ITW 2004 that helped precipitate the current line of research. They would also like to acknowledge the Associate Editor and an anonymous reviewer for their extensive and useful comments and suggestions during the revision process.

REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[2] C. Chang, "Streaming source coding with delay," Ph.D. dissertation, Dept. Electr. Eng. Comput. Sci., Univ. California, Berkeley, Berkeley, CA, USA, 2007.

[3] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 226–228, Mar. 1975.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

[5] I. Csiszàr and J. Kórner, *Information Theory, Coding Theorems for Discrete Memoryless Systems*. Budapest, Hungary: Akadémiai Kiadó, 1981.

[6] P. K. Dragotti and M. Gastpar, *Distributed Source Coding: Theory, Algorithms, and Applications*. San Diego, CA, USA: Academic Press, 2009.

[7] S. C. Draper, "Universal incremental Slepian-Wolf coding," in *Proc. Allerton Conf.*, Oct. 2004, pp. 1332–1341.

[8] S. C. Draper, C. Chang, and A. Sahai, "Sequential random binning for streaming distributed source coding," in *Proc. Int. Symp. Inf. Theory*, Sep. 2005, pp. 1396–1400.

[9] F. Dufaux, W. Gao, S. Tubaro, and A. Vetro, "Distributed video coding: Trends and perspectives," *J. Image Video Process.*, vol. 2009, pp. 508167-1–508167-3, Jan. 2009.

[10] A. W. Eckford and W. Yu, "Rateless Slepian-Wolf codes," in *Proc. 39th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2005, pp. 1757–1761.

[11] G. Forney, "Convolutional codes III. Sequential decoding," *Inf. Control*, vol. 25, no. 3, pp. 267–297, 1974.

[12] R. G. Gallager, "Source coding with side information and universal coding," Mass. Instit. Tech., Cambridge, MA, USA, Tech. Rep. LIDS-P-937, 1976.

[13] A. El Gamal and A. Orlitsky, "Interactive data compression," in *Proc. 25th Annu. Symp. Found. Comput. Sci.*, Oct. 1984, pp. 100–108.

[14] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[15] D.-K. He, L. A. Lastras-Montano, E.-H. Yang, A. Jagmohan, and J. Chen, "On the redundancy of Slepian-Wolf coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5607–5627, Dec. 2009.

[16] V. N. Koshelev, "On a problem of separate coding of two dependent sources," *Prob. Peredachi Inf.*, vol. 13, no. 1, pp. 26–32, 1977.

[17] P. Koulgi, E. Tuncel, S. Regunathan, and K. Rose, "On zero-error coding of correlated sources," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 2856–2873, Nov. 2003.

[18] J. Meng, E.-H. Yang, and Z. Zhang, "Tree interactive encoding and decoding: Conditionally ϕ-mixing sources," in *Proc. Int. Symp. Inf. Theory*, Aug. 2011, pp. 1–5.

[19] F. Pereira, C. Brites, J. Ascenso, and M. Tagliasacchi, "Wyner-Ziv video coding: A review of the early architectures and further developments," in *Proc. IEEE Int. Conf. Multimedia Exposit.*, Jun. 2008, pp. 625–628.

[20] R. Puri, S. S. Pradhan, and K. Ramchandran, "n-channel multiple descriptions: Theory and construction," in *Proc. Data Compress. Conf.*, Apr. 2002, pp. 262–271.

[21] S. Rajagopalan and L. Schulman, "A coding theorem for distributed computation," in *Proc. STOC (ACM Symp. Theory Comp.)*, 1994, pp. 790–799.

[22] A. Sahai, "Why block length and delay behave differently if feedback is present," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1860–1886, May 2008.

[23] A. Sahai and S. K. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link. Part I: Scalar systems," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.

[24] A. Sahai and S. K. Mitter, "Source coding and channel requirements for unstable processes," *IEEE Trans. Inf. Theory*, 2006.

[25] A. Sahai and H. Palaiyanur, "A simple encoding and decoding strategy for stabilization over discrete memoryless channels," in *Proc. Allerton Conf.*, Sep. 2005, pp. 538–547.

[26] N. Shulman and M. Feder, "Source broadcasting with an unknown amount of receiver side information," in *Proc. Inf. Theory Workshop*, Oct. 2002, pp. 127–130.

[27] 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.

[28] R. T. Sukhavasi and B. Hassibi, "Anytime reliable codes for stabilizing plants over erasure channels," in *Proc. Int. Conf. Control*, Dec. 2011, pp. 5249–5259.

[29] T. Weissman and A. El Gamal, "Source coding with limited-look-ahead side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5218–5239, Dec. 2006.

[30] L. Weng, S. S. Pradhan, and A. Anastasopoulos, "Error exponent regions for Gaussian broadcast and multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2919–2942, Jul. 2008.

[31] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[32] E.-H. Yang and D.-K. He, "Interactive encoding and decoding for one way learning: Near lossless recovery with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1808–1824, Apr. 2010.

Stark C. Draper (S'99–M'03) received the M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), and the B.S. and B.A. degrees in electrical engineering and history, respectively, from Stanford University.

He is an Associate Professor of Electrical and Computer Engineering at the University of Toronto, Canada. From 2007-2014 he was an Assistant Professor and an Associate Professor at the University of Wisconsin, Madison. Before moving to the University of Wisconsin he was with the Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA. He has held postdoctoral positions in the Wireless Foundations, University of California, Berkeley, and in the Information Processing Laboratory, University of Toronto. He has worked at Arraycomm, San Jose, CA, the C. S. Draper Laboratory, Cambridge, MA, and Ktaadn, Newton, MA. His research interests include communication and information theory, error-correction coding, statistical signal processing and optimization, security, and application of these disciplines to computer architecture.

Dr. Draper has received the NSF CAREER Award, the 2010 MERL President's Award, the UW ECE Gerald Holdridge Teaching Award, the MIT Carlton E. Tucker Teaching Award, an Intel Graduate Fellowship, Stanford's Frederick E. Terman Engineering Scholastic Award, and a U.S. State Department Fulbright Fellowship.

Cheng Chang received a B.E. degree from Tsinghua University, Beijing, in 2000, and a Ph.D. degree from University of California at Berkeley in 2007. He is currently a quantitative analyst with the D. E. Shaw Group in New York. In 2008, he spent a year in the information theory group at HP Labs as a postdoctoral researcher. His research interests are in signal processing, control theory, information theory and machine learning. He is the founder of www.24theory.com.

Anant Sahai (S'94–M'00) received the B.S. degree from the University of California (UC), Berkeley in 1994 and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996 and 2001, respectively. He joined the Department of Electrical Engineering and Computer Sciences at UC Berkeley in 2002 and is affiliated with the Wireless Foundations Center and the Berkeley Wireless Research Center. In 2001, he spent a year as a Research Scientist with the wireless startup Enuvis, developing adaptive algorithms for extremely sensitive GPS receivers implemented using software-defined radios. Prior to that, he was a graduate student at the Laboratory for Information and Decision Systems at MIT. His research interests are in wireless communication, signal processing, information theory, and distributed control. He is particularly interested in feedback, error exponents, and issues of spectrum sharing.