

# The price of ignorance: The impact of side-information on delay for lossless source-coding

Cheng Chang and Anant Sahai

## Abstract

Inspired by the context of compressing encrypted sources, this paper considers the general tradeoff between rate, end-to-end delay, and probability of error for lossless source coding with side-information. The notion of end-to-end delay is made precise by considering a sequential setting in which source symbols are revealed in real time and need to be reconstructed at the decoder within a certain *fixed* latency requirement. Upper bounds are derived on the reliability functions with delay when side-information is known only to the decoder as well as when it is also known at the encoder.

When the encoder is not ignorant of the side-information (including the trivial case when there is no side-information), it is possible to have substantially better tradeoffs between delay and probability of error at all rates. This shows that there is a fundamental *price of ignorance* in terms of end-to-end delay when the encoder is not aware of the side information. This effect is not visible if only fixed-block-length codes are considered. In this way, side-information in source-coding plays a role analogous to that of feedback in channel coding.

While the theorems in this paper are asymptotic in terms of long delays and low probabilities of error, an example is used to show that the qualitative effects described here are significant even at short and moderate delays.

## Index Terms

Real-time source coding, delay, reliability functions, error exponents, side-information, sequential coding, encryption

# The price of ignorance: The impact of side-information on delay for lossless source-coding

## I. INTRODUCTION

There are two surprising classical results pertaining to encoder “ignorance:” Shannon’s finding in [1] that the capacity of a memoryless channel is unchanged if the encoder has access to feedback and the Slepian-Wolf result in [2] that side-information at the encoder does not reduce the data-rate required for lossless compression. When the rate is not at the fundamental limit (capacity or conditional entropy), the error probability converges to zero exponentially in the allowed system delay — with block-length serving as the traditional proxy for delay in information theoretic studies. Dobrushin in [3] and Berlekamp in [4] followed up on Shannon’s result to show that feedback also does not improve<sup>1</sup> the block-coding error exponent in the high-rate regime (close to capacity) for symmetric channels. Similarly, Gallager in [6] and Csiszár and Körner in [7] showed that the block-coding error exponents for lossless source-coding also do not improve with encoder-side-information in the low rate regime (close to the conditional entropy). These results seemed to confirm the overall message that the advantages of encoder knowledge are limited to possible encoder/decoder implementation complexity reductions, not to anything more basic like rate or probability of error.

Once low complexity channel codes were developed that did not need feedback, mathematical and operational duality (See e.g. [8], [9]) enabled corresponding advances in low complexity distributed source codes. These codes then enabled radical new architectures for media coding in which the complexity could be shifted from the encoder to the decoder [10], [11]. Even more provocatively, [12] introduced a new architecture for information-theoretic secure communication illustrated as a shift from Figure 1 to Figure 2. By viewing Shannon’s one-time-pad from [13] as virtual side information, Johnson, *et al* in [12] showed that despite being marginally white and uniform, encrypted data could be compressed just as effectively by a system that does not have access to the key, as long as decoding takes place jointly with decryption. However, all of this work followed the traditional fixed-block-length perspective on source and channel coding.

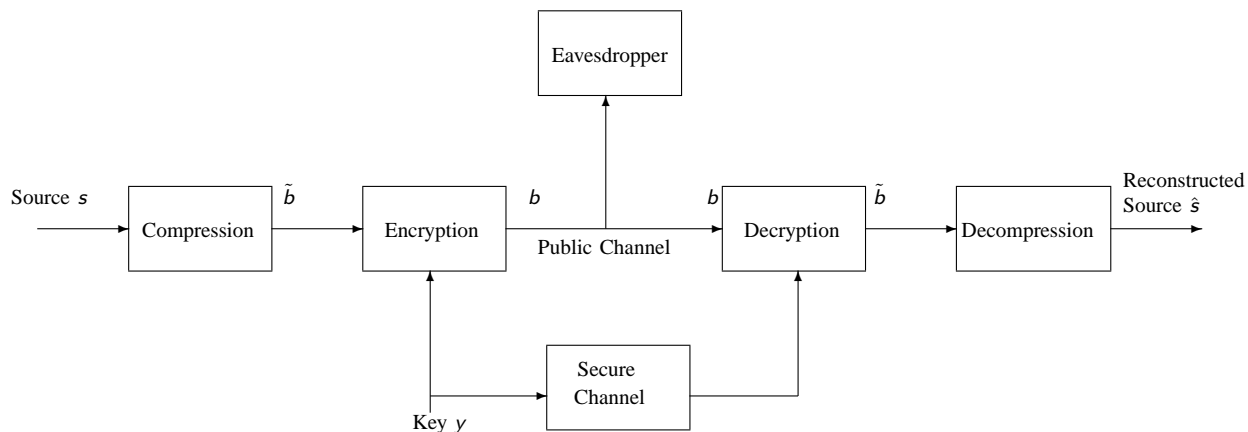


Fig. 1. The traditional compression/encryption system for sources with redundancy. (Figure adapted from [12])

Recently, it has become clear that the behavior of fixed-block-length codes and fixed-delay codes can be quite different in contexts where the message to be communicated is revealed to the encoder gradually as time progresses rather than being known all at once. In our entire discussion, the assumption is that information arises as a stream generated in real time at the source (e.g. voice, video, or sensor measurements) and it is useful to the destination in

<sup>1</sup>The history of feedback and its impact on channel reliability is reviewed in detail in [5].

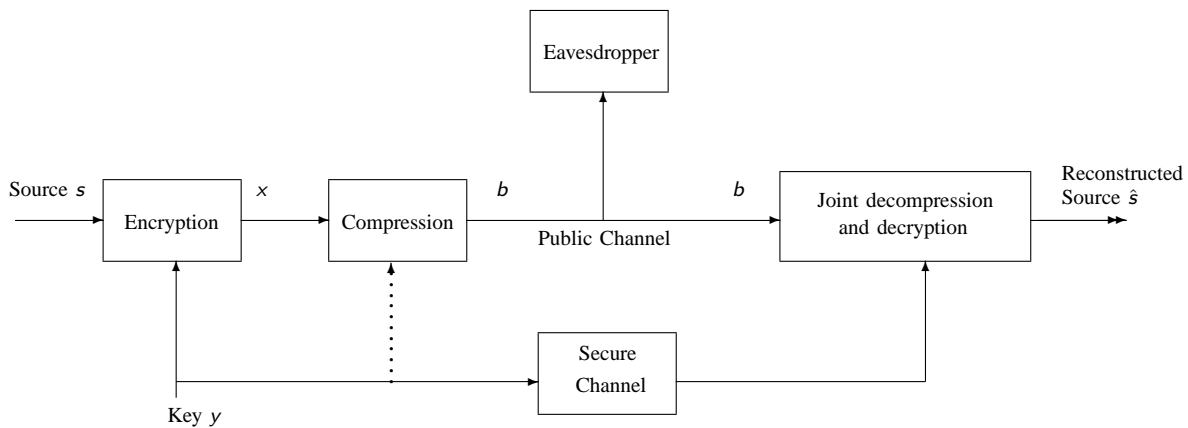


Fig. 2. The novel compression/encryption system in which a message is first encrypted and then compressed by the “ignorant” encoder. (Figure adapted from [12])

finely grained increments (e.g. a few milliseconds of voice, a single video frame, etc.). The encoded bitstream is also assumed to be transported at a steady rate. The acceptable end-to-end delay is determined by the application and can often be much larger than the natural granularity of the information being communicated (e.g. voice may tolerate a delay of hundreds of milliseconds despite being useful in increments of a few milliseconds). The end-to-end delay perspective here is common in the networking community. This is different from cases in which information arises in large bursts with each burst needing to be received by the destination before the next burst even becomes available at the source.

[5] shows that unlike the block channel coding reliability functions, the reliability function with respect to *fixed* end-to-end delay can in fact improve dramatically with feedback for essentially all DMCs at high rates.<sup>2</sup> The asymptotic factor reduction in end-to-end delay enabled by feedback approaches infinity as the message rate approaches capacity for generic DMCs. In addition, the nature of the dominant error events changes. Consider time relative to when a message symbol enters the encoder. Without feedback, errors are usually caused by *future* channel atypicality. When feedback is present, it is a *combination of past and future* atypicality that causes errors.

The results in [5] give a precise interpretation to the channel-coding half of Shannon’s intriguingly prophetic comment at the close of [16]:

“[the duality of source and channel coding] can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past and cannot control it; we may control the future but have no knowledge of it.”

One of the side benefits of this paper is to make Shannon’s comment similarly precise on the source coding side. Rather than worrying about what the appropriate granularity of information should be, the formal problem is specified at the individual source symbol level. If a symbol is not delivered correctly by its deadline, it is considered to be erroneous. The upper and lower bounds of this paper turn out to not depend on the choice of information granularity, only on the fact that the granularity is much finer than the tolerable end-to-end delay.

Here, we show that when decoder side-information is also available at the encoder, the dominant error event involves only the *past* atypicality of the source. This gives an upper bound on the fixed-delay error exponent that is the lossless source-coding counterpart to the “uncertainty focusing bound” given in [5] for channel coding with feedback. This bound is also shown to be asymptotically achievable at all rates. When side-information is present only at the decoder, [17] showed that the much slower random-coding error exponent is attainable with end-to-end delay. Here, an upper bound is given on the error exponent that matches the random-coding bound from [17] at low rates for appropriately symmetric cases — like the case of compressing encrypted data from [12]. This shows that there is a fundamental price of encoder ignorance that must be paid in terms of required end-to-end delay.

Section II fixes notation, gives the problem setup, and states the main results of this paper after reviewing the relevant classical results. Section III evaluates a specific numerical example to show the penalties of encoder

<sup>2</sup>It had long been known that the reliability function with respect to *average* block-length can improve [14], but there was a mistaken assertion by Pinsker in [15] that the fixed-delay exponents do not improve with feedback.

ignorance. It also demonstrates how the delay penalty continues to be substantial even in the non-asymptotic regime of short end-to-end delays and moderately small probability of error requirements. Section IV gives the proof for the fixed delay reliability function when both encoder and decoder have access to side-information. Section V proves the upper-bound on the fixed-delay reliability function when the encoder is ignorant of the side-information and the appendices show that it is tight for the symmetric case. Finally, Section VI gives some concluding remarks by pointing out the parallels between the source and channel coding stories.

## II. NOTATION, PROBLEM SETUP AND MAIN RESULTS

In this paper, all sources are iid random processes from finite alphabets where the finite alphabets are identified with the first few non-negative integers.  $x$  and  $y$  are random variables taking values in  $\mathcal{X}$  and  $\mathcal{Y}$ , with  $x$  and  $y$  used to denote realizations of the random variables. Without loss of generality, assume that  $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$ , the marginals  $p_x(x) > 0$  and  $p_y(y) > 0$ . The basic problem formulation is illustrated in Figure 3 for the cases with or without encoder access to the side-information.

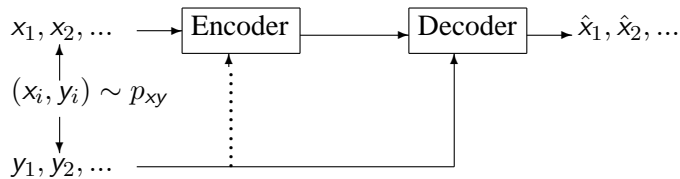


Fig. 3. Lossless source coding with encoder/decoder side-information.

The goal is to losslessly communicate the source  $x$ , drawn from a joint distribution  $p_{xy}$  on  $x, y$ , over a fixed rate bit-pipe. The decoder is always assumed to have access to the side-information  $y$ , and it may or may not be available to the encoder as well.

Rather than being known entirely in advance, the source symbols enter the encoder in a real-time fashion. (Illustrated in Figure 4) For convenience, time is counted in terms of source symbols: we assume that the source  $\mathcal{S}$  generates a pair of source symbols  $(x, y)$  per second from the finite alphabet  $\mathcal{X} \times \mathcal{Y}$ . The  $j$ 'th source symbol  $x_j$  is not known at the encoder until time  $j$  and similarly for  $y_j$  at the decoder (and possibly encoder). Rate  $R$  operation means that the encoder sends 1 binary bit to the decoder every  $\frac{1}{R}$  seconds. Throughout the paper the focus is on cases with  $H_{x|y} < R < \log_2 |\mathcal{X}|$ , since the lossless coding problem becomes trivial outside of that range.

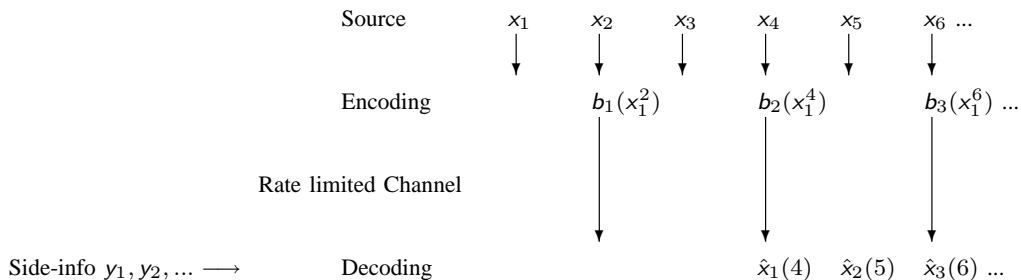


Fig. 4. Time line for fixed-delay source coding with decoder side-information: rate  $R = \frac{1}{2}$ , delay  $\Delta = 3$ .

*Definition 1:* A rate  $R$  encoder  $\mathcal{E}$  is a sequence of maps  $\{\mathcal{E}_j\}, j = 1, 2, \dots$ . The outputs of  $\mathcal{E}_j$  are the bits that are communicated from time  $j - 1$  to  $j$ . When the encoder does not have access to the decoder side-information:

$$\begin{aligned} \mathcal{E}_j : \mathcal{X}^j &\longrightarrow \{0, 1\}^{\lfloor jR \rfloor - \lfloor (j-1)R \rfloor} \\ \mathcal{E}_j(x_1^j) &= b_{\lfloor (j-1)R \rfloor + 1}^{\lfloor jR \rfloor} \end{aligned}$$

When the encoder does have access to the decoder side-information:

$$\begin{aligned}\mathcal{E}_j &: \mathcal{X}^j \times \mathcal{Y}^j \longrightarrow \{0, 1\}^{\lfloor jR \rfloor - \lfloor (j-1)R \rfloor} \\ \mathcal{E}_j(x_1^j, y_1^j) &= b_{\lfloor (j-1)R \rfloor + 1}^{\lfloor jR \rfloor}\end{aligned}$$

*Definition 2:* A fixed delay  $\Delta$  decoder  $\mathcal{D}^\Delta$  is a sequence of maps  $\{\mathcal{D}_j^\Delta\}$ ,  $j = 1, 2, \dots$ . The input to  $\mathcal{D}_j^\Delta$  are the all the bits emitted by the encoder until time  $j + \Delta$  as well as the side-information  $y_1^{j+\Delta}$ . The output is an estimate  $\hat{x}_j$  for the source symbol  $x_j$ .

Alternatively, a family of decoders indexed by different delays can be considered together. For these, the output is a list  $\hat{x}(j) = (\hat{x}_1(j), \dots, \hat{x}_j(j))$ .

$$\begin{aligned}\mathcal{D}_j^\Delta &: \{0, 1\}^{\lfloor jR \rfloor} \times \mathcal{Y}^j \longrightarrow \mathcal{X} \\ \mathcal{D}_j^\Delta(b_1^{\lfloor jR \rfloor}, y_1^j) &= \hat{x}_{j-\Delta}(j)\end{aligned}$$

where  $\hat{x}_{j-\Delta}(j)$  is the estimate of  $x_{j-\Delta}$  at time  $j$  and thus has an end-to-end delay of  $\Delta$  seconds.

The problem of lossless source-coding is considered by examining the asymptotic tradeoff between delay and the probability of symbol error:

*Definition 3:* A family (indexed by delay  $\Delta$ ) of rate  $R$  sequential source codes  $\{(\mathcal{E}^\Delta, \mathcal{D}^\Delta)\}$  achieves fixed-delay reliability  $E(R)$  if for all  $\epsilon > 0$ , there exists  $K < \infty$ , s.t.  $\forall i, \Delta > 0$

$$\Pr(x_i \neq \hat{x}_i(i + \Delta)) \leq K 2^{-\Delta(E(R) - \epsilon)}$$

when encoder  $\mathcal{E}^\Delta$  is used to do the encoding of the source and  $\mathcal{D}^\Delta$  is the decoder used to recover  $\hat{x}$ .

It is important to see that all source positions  $i$  require equal protection in terms of probability of error, but the probability of error can never be made uniform over the source realizations themselves since it is the source that is the main source of randomness in the problem!

#### A. Review of block source coding with side information

Before stating the new results, it is useful to review the classical fixed-block-length coding results. In fixed-block-length coding, the encoder has access to  $x_1^n$  all at once (as well as  $y_1^n$  if it has access to the side-information) and produces  $nR$  bits all at once. These bits go to the block decoder along with the side-information  $y_1^n$  and the decoder then produces estimates  $\hat{x}_1^n$  all at once. While the usual error probability considered is the block error probability  $\Pr(x_1^n \neq \hat{x}_1^n) = \Pr(x_1^n \neq \mathcal{D}_n(\mathcal{E}_n(x_1^n)))$ , there is no difference between the symbol error probability and the block-error probability on an exponential scale.

The relevant error exponents  $E(R)$  are considered in the limit of large block-lengths, rather than end-to-end delays.  $E(R)$  is achievable if  $\exists$  a family of  $\{(\mathcal{E}_n, \mathcal{D}_n)\}$ , s.t.

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log_2 \Pr(x_1^n \neq \hat{x}_1^n) = E(R) \quad (1)$$

The relevant results of [7], [6] are summarized into the following theorem.

*Theorem 1:* If the block-encoder does not have access to the side-information, the best possible block-error exponent is sandwiched between two bounds:  $E_{si,b}^l(R) \leq E_{si,b}(R) \leq E_{si,b}^u(R)$  where

$$E_{si,b}^l(R) = \min_{q_{xy}} \{D(q_{xy} \| p_{xy}) + \max\{0, R - H(q_{x|y})\}\} \quad (2)$$

$$= \sup_{0 \leq \rho \leq 1} \rho R - E_0(\rho) \quad (3)$$

$$E_{si,b}^u(R) = \min_{q_{xy}: H(q_{x|y}) \geq R} D(q_{xy} \| p_{xy}) \quad (4)$$

$$= \sup_{0 \leq \rho} \rho R - E_0(\rho) \quad (5)$$

where

$$E_0(\rho) = \log_2 \sum_y \left( \sum_x p_{xy}(x, y)^{\frac{1}{1+\rho}} \right)^{(1+\rho)} \quad (6)$$

is the Gallager function for the source with side-information.

The lower-bound corresponds to the performance of random-binning with MAP decoding. The upper and lower bounds agree for rates close to  $H(p_{x|y})$ , specifically  $R \leq \frac{\partial E_0(\rho)}{\partial \rho} \Big|_{\rho=1}$ .

If the encoder also has access to the side-information, then  $E_{si,b}^u(R)$  is the true error exponent at all rates since it can be achieved by simply encoding the conditional type of  $x_1^n$  given  $y_1^n$  and then encoding the index of the true realization within that conditional type.

If there is no side-information, then  $y = 0$  and the problem behaves like the case of side-information known at the encoder. (4) recovers the simple point-to-point fixed-block-length error exponent for lossless source coding. The resulting random and non-random error exponents are:

$$E_{s,b}^r(R, p_x) = \min_{q_x} \{D(q_x \| p_x) + \max\{0, R - H(q_x)\}\} \quad (7)$$

$$E_{s,b}(R, p_x) = \min_{q_x: H(q_x) \geq R} D(q_x \| p_x). \quad (8)$$

The Gallager function (6) in (5) and (3) also simplifies to

$$E_0(\rho) = (1 + \rho) \log_2 \left( \sum_x p_x(x)^{\frac{1}{1+\rho}} \right). \quad (9)$$

### B. Main results

[17] shows that the random coding bound  $E_{si,b}^l(R)$  is achievable with respect to end-to-end delay even without the encoder having access to the side-information. Thus, the factor of two increase in delay caused by using a fixed-block-length code in a real-time context is unnecessary. [17] uses a randomized sequential binning strategy with either MAP decoding or a universal decoding scheme that works for any iid source. [18] shows that the same asymptotic tradeoff with delay is achievable using a more computationally friendly stack-based decoding algorithm if the underlying joint distribution is known. However, it turns out that the end-to-end delay performance can be much better if the encoder has access to the side-information.

*Theorem 2:* For fixed rate  $R$  lossless source-coding of an iid source with side-information present at both the receiver and encoder, the asymptotic error exponent  $E_{ei}(R)$  with fixed end-to-end delay is given by the source uncertainty-focusing bound:

$$E_{ei}(R) = \inf_{\alpha > 0} \frac{1}{\alpha} E_{si,b}^u((\alpha + 1)R) \quad (10)$$

where  $E_{si,b}^u$  is defined in (4) and (5). The source uncertainty-focusing bound can also be expressed parametrically in terms of the Gallager function  $E_0(\rho)$  from (6):

$$\begin{aligned} E_{ei}(R) &= E_0(\rho) \\ R &= \frac{E_0(\rho)}{\rho} \end{aligned} \quad (11)$$

This bound generically approaches  $R = H(x|y)$  at strictly positive slope of  $2H(x|y) / \frac{\partial^2 E_0(0)}{\partial \eta^2}$ . When  $\frac{\partial^2 E_0(0)}{\partial \eta^2} = 0$ , the fixed-delay reliability function jumps discontinuously from zero to infinity.

Furthermore, this bound is asymptotically achievable by using universal fixed-to-variable block codes whose resulting data bits are smoothed to fixed-rate  $R$  through a FIFO queue with an infinite buffer size. This code is universal over iid sources as well as end-to-end delays that are sufficiently long (the block-length for the code is much smaller than the asymptotically large end-to-end delay constraint).

*Theorem 3:* For fixed rate  $R$  lossless source-coding of an iid source with side-information *only* at the receiver, the asymptotic error exponent  $E_{si}(R)$  with fixed end-to-end delay must satisfy  $E_{si}(R) \leq E_{si}^u(R)$ , where

$$\begin{aligned} E_{si}^u(R) = \min \{ & \inf_{q_{xy}, \alpha \geq 1: H(q_{x|y}) > (1+\alpha)R} \frac{1}{\alpha} D(q_{xy} \| p_{xy}), \\ & \inf_{q_{xy}, 1 \geq \alpha \geq 0: H(q_{x|y}) > (1+\alpha)R} \frac{1-\alpha}{\alpha} D(q_x \| p_x) + D(q_{xy} \| p_{xy}) \} \end{aligned} \quad (12)$$

For symmetric cases (such as those depicted in Figure 5), we have the following corollary:

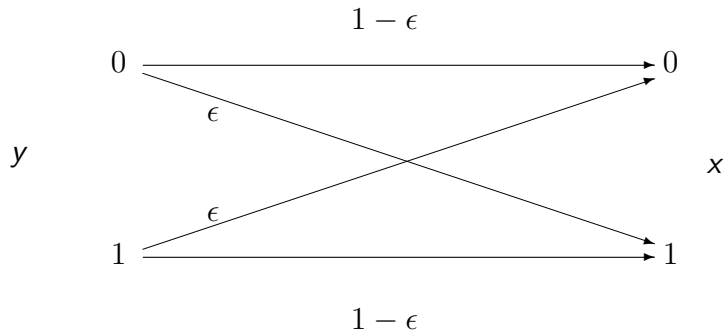


Fig. 5. A joint distribution on  $x, y$  that comes from a discrete memoryless channel connecting the two together and where the  $y$  is uniform and independent of the channel.

*Corollary 1:* Consider iid  $(x, y) \sim p_{xy}$  such that the side-information  $y$  is uniform on  $\mathcal{Y}$  and  $x = y \oplus s$ , where  $s \sim p_s$  is independent of  $y$ . Then the asymptotic error exponent  $E_{si}(R)$  with fixed delay must satisfy  $E_{si}(R) \leq E_{si,b}^u(R) = E_{s,b}(R, p_s)$  from (4) and (8).

Since [17] shows that  $E_{si,b}^u(R)$  is achievable at low rates, Corollary 1 is tight there.

### III. APPLICATION AND NUMERIC EXAMPLE

While the above results are general, they can be applied to the specific context of the [12] approach of compressing encrypted data. The general problem is depicted in Figure 6 in terms of joint encryption and compression. The goal is to communicate from end-to-end using a reliable fixed-rate bit-pipe in such a way that:

- The required rate of the bit-pipe is low.
- The probability of error is low for each source symbol.
- The end-to-end delay is small.
- Nothing is revealed to an eavesdropper that has access to the bitstream.

The idea is to find a good tradeoff among the first three while preserving the fourth. To support these goals, assume access to an infinite supply of common-randomness shared among the encoder and decoder that is not available to the eavesdropper. This can be used as a secret key. We are not concerned here with the size of the secret key.

This section evaluates the fixed-delay performance for both the compression-first and encryption-first systems as a way of showing the delay price of the encoder's ignorance of the side-information in the encryption-first approach. Nonasymptotic behavior is explored using a specific code for short values of end-to-end delay to verify that the price of ignorance also hits when delays are small.

#### A. Encryption/compression of streaming data: asymptotic results

The main results of this paper can be used to evaluate two candidate architectures: the traditional compression-first approach depicted in Figure 1 and the novel encryption-first approach proposed in [12] and depicted in Figure 2.

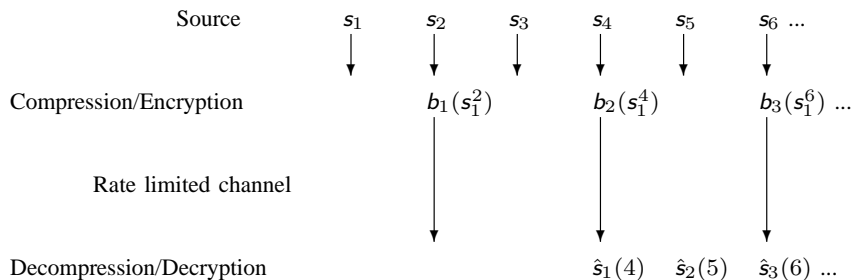


Fig. 6. Joint encryption and compression of streaming data with a fixed end-to-end delay constraint. Here the rate  $R = \frac{1}{2}$  bits per source symbol and delay  $\Delta = 3$ .



In practical terms, this means that if both the end-to-end delay and acceptable probability of symbol error are constrained by the application, then the approach of encryption followed by compression can end up requiring higher-rate bit-pipes.

### B. Numeric example including nonasymptotic results

Consider a simple source  $s$  with alphabet size 3,  $\mathcal{S} = \{A, B, C\}$  and distribution

$$p_s(A) = a \quad p_s(B) = \frac{1-a}{2} \quad p_s(C) = \frac{1-a}{2}$$

where  $a = 0.65$  for the plots and numeric comparisons.

1) *Asymptotic error exponents*: The different error exponents for fixed-block-length and fixed-delay source coding predict the asymptotic performance of different source coding systems when the end-to-end delay is long. We plot the source uncertainty-focusing bound  $E_{ei}(R)$ , the fixed-block-length error exponent  $E_{s,b}(R, p_s)$  and the random coding bound  $E_{s,b}^r(R, p_s)$  in Figure 9. For this source, the random coding and fixed-block-length error exponents are the same for  $R \leq \frac{\partial E_0(\rho)}{\partial \rho} \Big|_{\rho=1} = 1.509$ . Theorem 1 and Theorem 2 reveal that these error exponents govern the asymptotic performance of fixed-delay systems with and without encoder side-information.

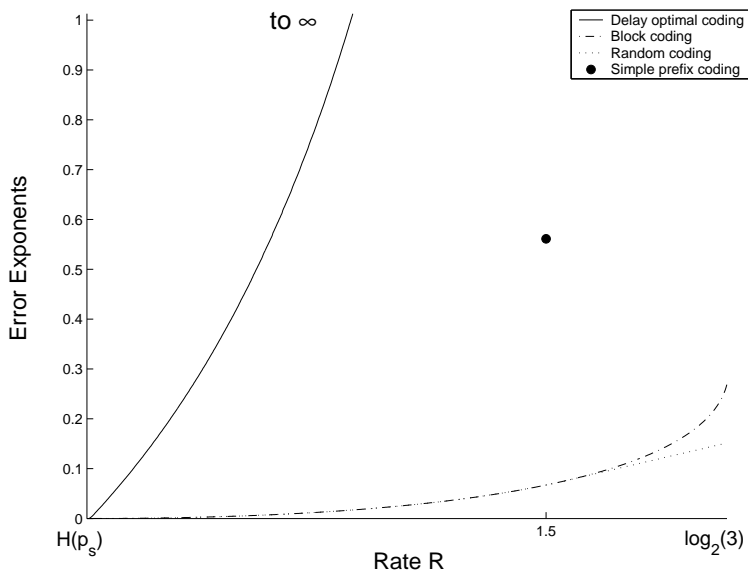


Fig. 9. Different source coding error exponents: fixed-delay error exponent  $E_s(R)$  with encoder side-information, fixed-block-length error exponent  $E_{s,b}(R, p_s)$ , and the random coding bound  $E_{s,b}^r(R, p_s)$ . The fixed-block-length bound also bounds the fixed-delay case without encoder side-information since the example here is symmetric.

Figure 10 plots the ratio of the source uncertainty-focusing bound over the fixed-block-length error exponent. The ratio tells asymptotically how many times longer the delay must be for the system built around an encoder that does not have access to the side-information. The smallest ratio is around 52 at a rate around 1.45.

2) *Non-asymptotic results*: The price of ignorance is so high, that even non-optimal codes with encoder side-information can outperform optimal codes without it. This section uses a very simple fixed-delay coding scheme using a prefix-free fixed-to-variable code[19] instead of the asymptotically optimal universal code described in Theorem 2. The input block-length is two, and the encoder uses the side-information to recover  $s$  before encoding it as:

$$\begin{aligned} AA &\rightarrow 0 \\ AB &\rightarrow 1000 \quad AC \rightarrow 1001 \quad BA \rightarrow 1010 \quad BB \rightarrow 1011 \\ BC &\rightarrow 1100 \quad CA \rightarrow 1101 \quad CB \rightarrow 1110 \quad CC \rightarrow 1111 \end{aligned}$$

For ease of analysis, the system is run at  $R = \frac{3}{2} < \frac{\partial E_0(\rho)}{\partial \rho} \Big|_{\rho=1} = 1.509$ . This means that the source generates 1 symbol per second and 3 bits are sent through the error-free bit-pipe every 2 seconds. The variable-rate of the

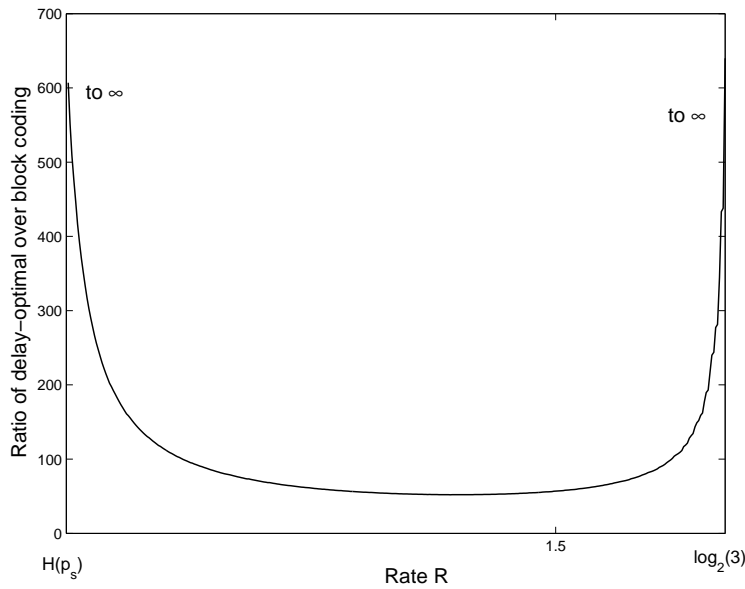


Fig. 10. Ratio of the fixed-delay error exponent with encoder side-information  $E_{ei}(R)$  over the fixed-block-length error exponent  $E_{s,b}(R, p_s)$ . This reflects the asymptotic factor increase in end-to-end delay required to compensate for the encoder being ignorant of the side-information available at the decoder.

code is smoothed through a FIFO queue with an infinite buffer in a manner similar to the buffer-overflow problem studied in [20], [21]. The entire coding system is illustrated in Figure 11.

It is convenient to examine time in increments of two seconds. The length of the codeword generated is either 1 or 4. The buffer is drained out by 3 bits per 2 seconds. Let  $L_k$  be the number of bits in the buffer as at time  $2k$ . Every two seconds, the number of bits  $L_k$  in the buffer either goes down by 2 if  $s_{2k-1}, s_{2k} = AA$  or goes up by 1 if  $s_{2k-1}s_{2k} \neq AA$ . If the queue is empty, the encoder can send arbitrary bits through the bit-pipe without causing confusion at the decoder because the decoder knows that the source only generates 1 source symbol per second and that it is caught up.

Source	AA	AB	BA	CC	AA	CA	AA	AA	AA	CB	AA	CC		
Prefix code	0	1000	1010	1111	0	1101	0	0	0	1110	0	1111		
Buffer	/	/	0	10	111	0	01	/	/	/	0	/	1	
Rate R bit-stream	***	0**	100	010	101	111	011	010	0**	0**	111	00*	111	
Decision		AA		AB	BA	CC	AA		CA	AA	AA		CB	AA

Fig. 11. Suboptimal prefix coding system in action. / indicates empty queue, \* indicates meaningless filler bits.

Clearly  $L_k, k = 1, 2, \dots$  forms a Markov chain with following transition matrix:  $L_k = L_{k-1} + 1$  with probability  $1 - a^2$ ,  $L_k = L_{k-1} - 2$  with probability  $a^2$ . The state transition graph is illustrated in Figure 12. For this Markov

chain, the stationary distribution can be readily calculated<sup>3</sup> [22].

$$\pi_k = Z \left( \frac{-1 + \sqrt{1 + \frac{4(1-q)}{q}}}{2} \right)^k \quad (13)$$

Where  $q = a^2$  and  $Z = 1 - \frac{-1 + \sqrt{1 + \frac{4(1-q)}{q}}}{2}$  is the normalization constant. For this example  $Z = 0.228$ . Notice that  $\pi_k$  is geometric and the stationary distribution exists as long as  $4 \frac{1-q}{q} < 8$  or equivalently  $q > \frac{1}{3}$ . In this example,  $a = 0.65$  and thus  $q = a^2 = 0.4225 > \frac{1}{3}$ , so the stationary distribution  $\pi_k$  exists.

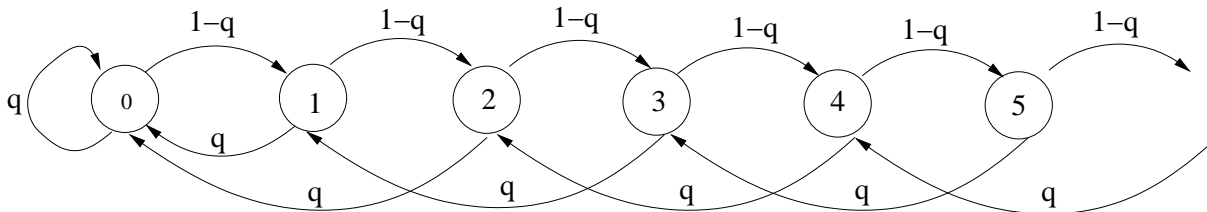


Fig. 12. Transition graph of a reflecting random walk  $L_k$  for queue length given the specified prefix-free code and the ternary source distribution  $\{a, \frac{1-a}{2}, \frac{1-a}{2}\}$  and fixed rate  $\frac{3}{2}$  bits per source symbol.  $q = a^2$  denotes the probability that  $L_k$  decrements by 2.

Assume  $\Delta$  is odd for convenience. For the above simple coding system, a decoding error can only happen if at time  $2k - 1 + \Delta$ , at least one bit of the codeword describing  $s_{2k-1}, s_{2k}$  is still in the queue. Since the queue is FIFO, this implies that there were too many bits awaiting transmission at time  $2k$  itself — ie that the number of bits  $L_k$  in the buffer at time  $2k$ , is larger than

$$\lfloor \frac{3}{2}(\Delta - 1) \rfloor - l(s_{2k-1}, s_{2k})$$

where  $l(s_{2k-1}, s_{2k})$  is the length of the codeword for  $s_{2k-1}, s_{2k}$ .  $l$  is 1 with probability  $q = a^2$  and 4 with probability  $1 - q = 1 - a^2$ . Notice that the length of the codeword for  $s_{2k-1}, s_{2k}$  is independent of  $L_k$  since the source symbols are iid. This gives the following upper bound on the error probability of decoding with delay  $\Delta$  when the system is in steady state<sup>4</sup>:

$$\begin{aligned} \Pr(\hat{s}_{2k}(2k - 1 + \Delta) \neq s_{2k}) &= \\ \Pr(\hat{s}_{2k-1}(2k - 1 + \Delta) \neq s_{2k-1}) &\leq \Pr(l(s_{2k-1}, s_{2k}) = 1) \Pr(L_k > \lfloor \frac{3}{2}(\Delta - 1) \rfloor - 1) \\ &\quad \Pr(l(s_{2k-1}, s_{2k}) = 4) \Pr(L_k > \lfloor \frac{3}{2}(\Delta - 1) \rfloor - 4) \\ &= q \sum_{j=\lfloor \frac{3}{2}(\Delta-1) \rfloor}^{\infty} \pi_j + (1-q) \sum_{j=\lfloor \frac{3}{2}(\Delta-1) \rfloor - 3}^{\infty} \pi_j \\ &= G \left( \frac{-1 + \sqrt{1 + \frac{4(1-q)}{q}}}{2} \right)^{\lfloor \frac{3}{2}(\Delta-1) \rfloor - 3} \end{aligned}$$

where  $G$  is the normalization constant

$$G = Z \left( q \sum_{j=0}^2 \left( \frac{-1 + \sqrt{1 + \frac{4(1-q)}{q}}}{2} \right)^j + (1-q) \right).$$

For this example,  $G = 0.360$ . Thus, the fixed-delay error exponent for this coding system is

$$\frac{3}{2} \log_2 \left( \frac{-1 + \sqrt{1 + \frac{4(1-q)}{q}}}{2} \right).$$

<sup>3</sup>The polynomial corresponding to the recurrence relation for the stationary distribution has three roots. One of them is 1 and the other is unstable since it has magnitude larger than 1. That leaves only one possibility for the stationary distribution.

<sup>4</sup>If the system is initialized to start in the zero state, then this bound remains valid since the system approaches steady state from below.

Figure 13 compares three different coding schemes in the non-asymptotic regime of short delays and moderate probabilities of error at rate  $\frac{3}{2}$ . As shown in Figure 9, at this rate the random coding error exponent  $E_{s,b}^r(R, p_s)$  is the same as the fixed-block-length error exponent  $E_{s,b}(R, p_s)$ . The block coding curve plotted is for an optimal coding scheme in which the encoder first buffers up  $\frac{\Delta}{2}$  symbols, encodes them into a length  $\frac{\Delta}{2}R$ -length binary sequence and uses the next  $\frac{\Delta}{2}$  seconds to transmit the message. This coding scheme gives an error exponent  $\frac{E_{s,b}(R, p_s)}{2}$  with delay in the limit of long delays.

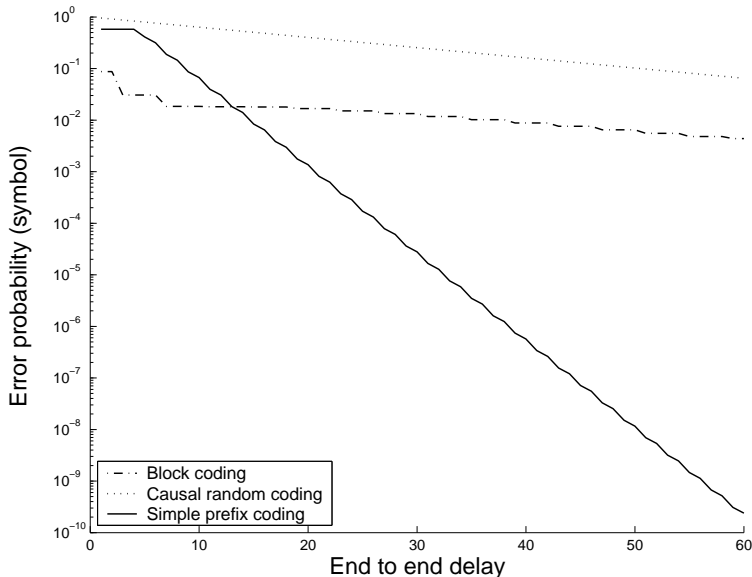


Fig. 13. Error probability vs delay (non-asymptotic results) illustrating the price of encoder ignorance.

The slope of these curves in Figure 13 indicates the error exponent governing how fast the error probability goes to zero with delay. Although smaller than the delay optimal error exponent  $E_s(R)$ , this simple coding strategy has a much higher fixed-delay error exponent than both sequential random coding and optimal *simplex* block coding. A simple calculation reveals that in order to get a  $10^{-6}$  symbol error probability, the delay requirement for our simple scheme is  $\sim 40$ , for causal random coding is around  $\sim 303$ , and for optimal block coding is around  $\sim 374$ . Thus, the price of encoder ignorance is very significant even in the non-asymptotic regime and fixed-block-length codes are very suboptimal from an end-to-end delay point of view.

#### IV. ENCODERS WITH SIDE-INFORMATION

The goal of this section is to prove Theorem 2 directly.

##### A. Achievability

The achievability of  $E_{ei}(R)$  is shown using a simple fixed-to-variable<sup>6</sup> length universal code that has its output rate smoothed through a FIFO queue. Because the end-to-end delay experienced by a symbol is dominated by the time spent waiting in the queue, and the queue is drained at a deterministic rate, the end-to-end delay experienced by a symbol is essentially proportional to the length of the queue when that symbol arrives. Thus on the achievability side, Theorem 2 can be viewed as a corollary to results on the buffer-overflow exponent for fixed-to-variable length codes. The buffer-overflow exponent was first derived in [20] for cases without any side-information at all. Here, we simply state the coding strategy used and leave the detailed analysis for Appendix A.

The strategy only depends on the size of the source alphabets  $|\mathcal{X}|, |\mathcal{Y}|$ , not on the distribution of the source.

<sup>5</sup>We ran a linear regression on the data  $y_\Delta = \log_{10} P_e(\Delta)$ ,  $x_\Delta = \Delta$  as shown in Figure 9 from  $\Delta = 80$  to  $\Delta = 100$  to extrapolate the  $\Delta$ , s.t.  $\log_{10} P_e(\Delta) = -6$ .

<sup>6</sup>Fixed-to-variable was chosen for ease of analysis. It is likely that variable-to-fixed and variable-to-variable length codes can also be used as the basis for an optimal fixed-delay source coding system.

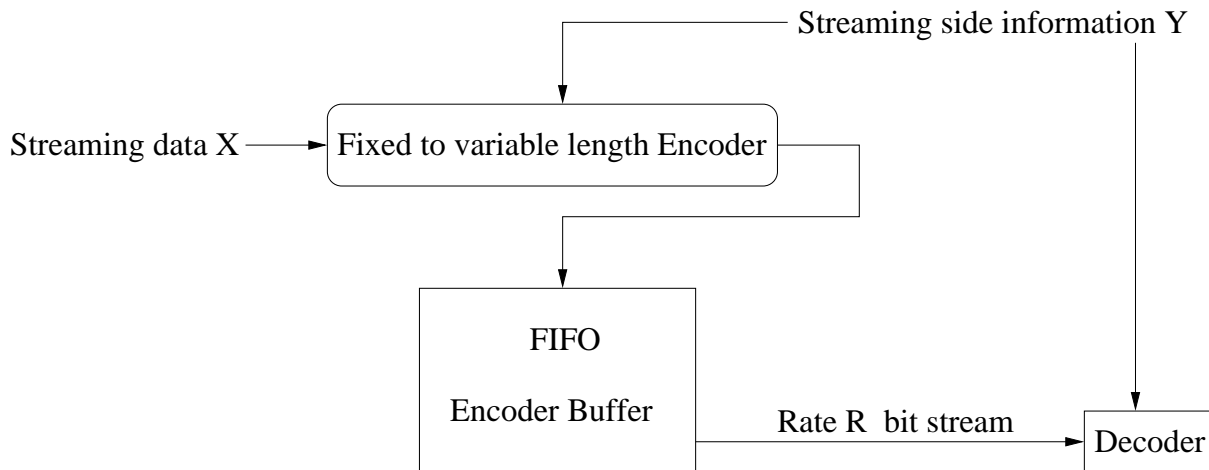


Fig. 14. A universal fixed-delay lossless source coding system built around a fixed-to-variable block-length code.

First, a finite block-length  $N$  is chosen that is much smaller than the asymptotically large target end-to-end delay  $\Delta$ . For a discrete memoryless source  $x$ , side information  $y$  and large block-length  $N$ , an optimal fixed-to-variable code is given in [7] and consists of three stages:

- 1) Start with a 1.
- 2) Describe the joint type of the block  $\vec{x}_i$  (the  $i$ 'th block of length  $N$ ) and  $\vec{y}_i$ . This costs at most a fixed  $1 + |\mathcal{X}||\mathcal{Y}|\log_2 N$  bits per block.
- 3) Describe which particular realization has occurred for  $\vec{x}_i$  by using a variable  $NH(\vec{x}_i|\vec{y}_i)$  bits where  $H(\vec{x}_i|\vec{y}_i)$  is the empirical conditional entropy of sequence  $\vec{x}_i$  given  $\vec{y}_i$ .

This code is obviously prefix-free. When the queue is empty, the fixed-rate  $R$  encoder can send a 0 without introducing any ambiguity. The total end-to-end delay experienced by any individual source-symbol is then upper-bounded by  $N$  (how long it must wait to be assembled into a block) plus  $\frac{1}{R}$  times the length of the queue once it has been encoded.

Write  $l(\vec{x}_i, \vec{y}_i)$  as the random total length of the codeword for  $\vec{x}_i, \vec{y}_i$ . Then

$$NH(\vec{x}_i|\vec{y}_i) \leq l(\vec{x}_i, \vec{y}_i) = N(H(\vec{x}_i|\vec{y}_i) + \epsilon_N) \quad (14)$$

where  $\epsilon_N \leq \frac{2+|\mathcal{X}||\mathcal{Y}|\log_2(N+1)}{N}$  goes to 0 as  $N$  gets large.

Because the source is iid, the lengths of the blocks are also iid. Each one has a length whose distribution can be bounded using Theorem 1. From there, there are two paths to show the desired result. One path uses Corollary 6.1 of [5] and for that, all that is required is a lemma parallel to Lemma 7.1 of [5] asserting that the length of the block has a distribution upperbounded by a constant plus a geometric random variable. Such a bound easily follows from the (5) formulation for the block-reliability function. We take a second approach proceeding directly using standard large deviations techniques. The following lemma bounds the probability of atypical source behavior for the sum of lengths.

*Lemma 1:* for all  $\epsilon > 0$ , there exists a block length  $N$  large enough so that there exists  $K < \infty$  such that for all  $n > 0$  and all  $H(x|y) < r < \log_2 |\mathcal{X}|$

$$\Pr\left(\sum_{i=1}^n l(\vec{x}_i, \vec{y}_i) > nNr\right) \leq K2^{-nN(E_{s_i,b}^u(r)-\epsilon)}. \quad (15)$$

*Proof:* : See Appendix .

At time  $(t + \Delta)N$ , the decoder *cannot* decode  $\vec{x}_t$  with 0 error probability iff the binary strings describing  $\vec{x}_t$  are *not* all out of the buffer yet. Since the encoding buffer is FIFO, this means that the number of outgoing bits from some time  $t_1$  to  $(t + \Delta)N$  is less than the number of the bits in the buffer at time  $t_1$  plus the number of incoming bits from time  $t_1$  to time  $tN$ . Suppose the buffer were last empty at time  $t_1 = tN - nN$  where  $0 \leq n \leq t$ . Given this, a decoding error could occur only if  $\sum_{i=0}^{n-1} l(\vec{x}_{t-i}, \vec{y}_{t-i}) > (n + \Delta)NR$ .

Denote the longest code length by  $l_{max} \leq 2 + |\mathcal{X}||\mathcal{Y}| \log_2(N+1) + N \log_2 |\mathcal{X}|$ . Then  $\Pr(\sum_{i=0}^{n-1} l(\vec{x}_{t-i}, \vec{y}_{t-i}) > (n+\Delta)NR) > 0$  only if  $n > \frac{(n+\Delta)NR}{l_{max}} > \frac{\Delta NR}{l_{max}} \triangleq \beta\Delta$ . So

$$\begin{aligned}
\Pr(\vec{x}_t \neq \vec{x}_t((t+\Delta)N)) &\leq \sum_{n=\beta\Delta}^t \Pr\left(\sum_{i=0}^{n-1} l(\vec{x}_{t-i}, \vec{y}_{t-i}) > (n+\Delta)NR\right) \\
&\stackrel{(a)}{\leq} \sum_{n=\beta\Delta}^t K_1 2^{-nN(E_{si,b}^u(\frac{(n+\Delta)NR}{nN})-\epsilon_1)} \\
&\stackrel{(b)}{\leq} \sum_{n=\gamma\Delta}^{\infty} K_2 2^{-nN(E_{si,b}^u(R)-\epsilon_2)} + \sum_{n=\beta\Delta}^{\gamma\Delta} K_2 2^{-\Delta N(\min_{\alpha>1} \{\frac{E_{si,b}^u(\alpha R)}{\alpha-1}\}-\epsilon_2)} \\
&\stackrel{(c)}{\leq} K_3 2^{-\gamma\Delta N(E_{si,b}^u(R)-\epsilon_2)} + |\gamma\Delta - \beta\Delta| K_3 2^{-\Delta N(E_{ei}(R)-\epsilon_2)} \\
&\stackrel{(d)}{\leq} K 2^{-\Delta N(E_{ei}(R)-\epsilon)}
\end{aligned} \tag{16}$$

where the large  $K_i$ 's and arbitrarily tiny  $\epsilon_i$ 's are properly chosen real numbers. (a) is true because of Lemma 1. Letting  $\gamma = \frac{E_{ei}(R)}{E_{si,b}^u(R)}$  in the first part of (b), we only need the fact that  $E_{si,b}^u(R)$  is non-decreasing with  $R$ . In the second part of (b), let  $\alpha = \frac{n+\Delta}{n}$  and choose the  $\alpha$  to minimize the error exponents. The first term of (c) comes from the sum of a geometric series. The second term of (c) follows from the definition of  $E_{ei}(R)$  in (10). (d) follows from the definition of  $\gamma$  above and by absorbing the linear term into the  $\epsilon$  in the exponent. ■

## B. Converse

The idea is to bound the best possible error exponent with fixed delay, without making any assumptions on the implementation of the encoder and decoder beyond the fixed end-to-end delay constraint. In particular, no assumption is made that the encoder works by encoding source symbols in small groups and then uses a queue to smooth out the rate. Instead, an encoder/decoder pair is considered that uses the fixed-delay system to construct a fixed-block-length system. The block-coding bounds of [7] are thereby translated to the fixed delay context. The arguments are analogous to the ‘‘uncertainty-focusing bound’’ derivation in [5] for the case of channel coding with feedback and the techniques originate in the convolutional code literature [23].

*Proof:* For simplicity of exposition, we ignore integer effects arising from the finite nature of  $\Delta$  and  $R$  etc. For every  $\alpha > 0$  and delay  $\Delta$ , consider a code running at fixed-rate till time  $\frac{\Delta}{\alpha} + \Delta$ . By this time, the decoder has committed to estimates for the source symbols up to time  $i = \frac{\Delta}{\alpha}$ . The total number of bits generated by the encoder during this period is  $(\frac{\Delta}{\alpha} + \Delta)R$ .

Now, relax the causality constraint at the encoder by giving it access to the first  $i$  source symbols all at once at the beginning of time, rather than forcing the encoder to get the source symbols gradually. Simultaneously, loosen the deadlines at the decoder to only demand correct estimates for the first  $i$  source symbols by the time  $\frac{\Delta}{\alpha} + \Delta$ . In effect, the deadline for decoding the *past* source symbols is extended to the deadline of the  $i$ -th symbol itself.

Any lower-bound to the symbol error probability of the new problem is clearly also a bound for the original problem. The difference between block error probability and symbol error probability is at most a factor of  $\frac{1}{i}$  and is insignificant on the exponential scale. Furthermore, the new problem is just a fixed-block-length source coding problem requiring the encoding of  $i$  source symbols into  $(\frac{\Delta}{\alpha} + \Delta)R$  bits. The rate per symbol is

$$\begin{aligned}
((\frac{\Delta}{\alpha} + \Delta)R) \frac{1}{i} &= ((\frac{\Delta}{\alpha} + \Delta)R) \frac{\alpha}{\Delta} \\
&= (\alpha + 1)R.
\end{aligned}$$

Theorem 2.15 in [7] tells us that such a code has a probability of error that is at least exponential in  $iE_{ei,b}((\alpha+1)R)$ . Since  $i = \frac{\Delta}{\alpha}$ , this translates into an error exponent of at most  $\frac{E_{ei,b}((\alpha+1)R)}{\alpha}$  with parameter  $\Delta$ .

Since this is true for all  $\alpha > 0$ , we have the uncertainty-focusing bound on the reliability function  $E_{ei}(R)$  with fixed delay  $\Delta$ :

$$E_{ei}(R) \leq \inf_{\alpha>0} \frac{1}{\alpha} E_{ei,b}((\alpha+1)R) \tag{17}$$

The minimizing  $\alpha$  tells how much of the past ( $\frac{\Delta}{\alpha}$ ) is involved in the dominant error event.

The source uncertainty-focusing bound can be expressed parametrically in terms of the Gallager function  $E_0(\rho)$  from (6) and its slope computed in the vicinity of the conditional entropy. This is shown in Appendix B. ■

## V. NO SIDE-INFORMATION AT THE ENCODER

This section proves the upper bound given by Theorem 3 for the fixed-delay error exponent for source coding without encoder side-information. This bound is valid for any generic joint distribution  $p_{xy}$ . The results are specialized to the symmetric case in Corollary 1, proved in Appendix H.

In the following analysis, it is conceptually useful to factor the joint probability to treat the source as a random variable  $x$  and consider the side-information  $y$  as the output of a discrete memoryless channel (DMC)  $p_{y|x}$  with  $x$  as input. This model is shown in Figure 15.

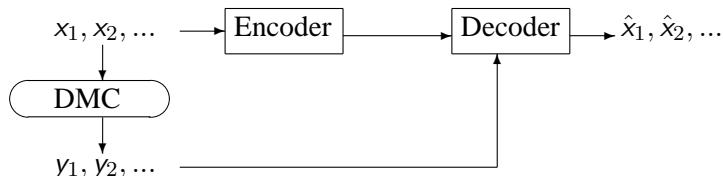


Fig. 15. Lossless source coding with side-information only at the decoder.

The theorem is proved using a variation of the bounding technique used in [5] (and originating in [15]) for the fixed-delay channel coding problem. Lemmas 2-7 are the source coding counterparts to Lemmas 4.1-4.4 in [5]. The idea of the proof is to assume a more powerful source decoder that has access to the previous source symbols (considered as feed-forward information) in addition to the encoded bits and the side-information. The second step is to construct a fixed-block-length source-coding scheme from the encoder and optimal feed-forward decoder. The third step is to show that if the side-information behaves atypically enough, then the decoding error probability will be large for many of the source symbols. The fourth step is to show that it is only future atypicality of the side-information that matters. This is because the feed-forward information allows the decoder to safely ignore all side-information concerning the source symbols that it already knows perfectly. The last step is to lower bound the probability of the atypical behavior and upper bound the error exponents. The proof spans the next several subsections.

### 1) Feed-forward decoders :

*Definition 4:* A delay  $\Delta$  rate  $R$  decoder  $\mathcal{D}^{\Delta,R}$  with feed-forward is a decoder  $\mathcal{D}_j^{\Delta,R}$  that also has access to the past source symbols  $x_1^{j-1}$  in addition to the encoded bits  $b_1^{\lfloor(j+\Delta)R\rfloor}$  and side-information  $y_1^{j+\Delta}$ .

Using this feed-forward decoder, the estimate of  $x_j$  at time  $j + \Delta$  is :

$$\hat{x}_j(j + \Delta) = \mathcal{D}_j^{\Delta,R}(b_1^{\lfloor(j+\Delta)R\rfloor}, y_1^{j+\Delta}, x_1^{j-1}) \quad (18)$$

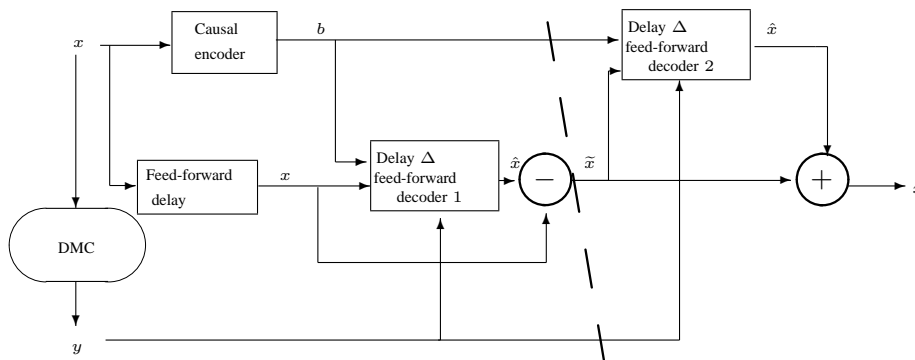


Fig. 16. A cutset illustration of the Markov Chain  $x_1^n - (\hat{x}_1^n, b_1^{\lfloor(n+\Delta)R\rfloor}, y_1^{n+\Delta}) - x_1^n$ . Decoder 1 and decoder 2 are type I and II delay  $\Delta$  rate  $R$  feed-forward decoders respectively.

*Lemma 2:* For any rate  $R$  encoder  $\mathcal{E}$ , the optimal delay  $\Delta$  rate  $R$  decoder  $\mathcal{D}^{\Delta,R}$  with feed-forward only needs to depend on  $b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}, x_1^{j-1}$

*Proof:* The source and side-information pair  $(x_i, y_i)$  is an iid random process and the encoded bits  $b_1^{\lfloor(j+\Delta)R\rfloor}$  are causal functions of  $x_1^{j+\Delta}$ . It is easy to see that the Markov chain  $y_1^{j-1} - (x_1^{j-1}, b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}) - x_j^{j+\Delta}$  holds since

$$\begin{aligned} & \Pr(x_j^{j+\Delta}, y_1^{j-1}, x_1^{j-1}, b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}) \\ &= \Pr(x_1^{j-1}, b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}) \Pr(x_j^{j+\Delta} | x_1^{j-1}, b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}) \Pr(y_1^{j-1} | x_1^{j-1}, b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}, x_j^{j+\Delta}) \\ &= \Pr(x_1^{j-1}, b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}) \Pr(x_j^{j+\Delta} | x_1^{j-1}, b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}) \Pr(y_1^{j-1} | x_1^{j-1}) \end{aligned}$$

Thus, conditioned on the past source symbols, the past side-information is completely irrelevant for optimal MAP estimation of  $x_j$ .  $\square$

Write the error sequence of the feed-forward decoder as  $\tilde{x}_i = x_i - \hat{x}_i$  by identifying the finite source alphabet with the appropriate finite group. Then we have the following property for the feed-forward decoders.

*Lemma 3:* Given a rate  $R$  encoder  $\mathcal{E}$ , the optimal delay  $\Delta$  rate  $R$  decoder  $\mathcal{D}^{\Delta,R}$  with feed-forward for symbol  $j$  only needs to depend on  $b_1^{\lfloor(j+\Delta)R\rfloor}, y_1^{j+\Delta}, \tilde{x}_1^{j-1}$

*Proof:* Proceed by induction. It holds for  $j = 1$  since there are no prior source symbols. Suppose that it holds for all  $j < k$  and consider  $j = k$ . By the induction hypothesis, the action of all the prior decoders  $j$  can be simulated using  $(b_1^{\lfloor(j+\Delta)R\rfloor}, y_1^{j+\Delta}, \tilde{x}_1^{j-1})$  giving  $\hat{x}_1^{k-1}$ . This in turn allows the recovery of  $x_1^{k-1}$  since we also know  $\tilde{x}_1^{k-1}$ . Thus the optimal feed-forward decoder can be expressed in this form.  $\square$

We call the feed-forward decoders in Lemmas 2 and 3 type I and II delay  $\Delta$  rate  $R$  feed-forward decoders respectively. Lemmas 2 and 3 tell us that feed-forward decoders can be thought in three ways: having access to all encoded bits, all side information and all past source symbols,  $(b_1^{\lfloor(j+\Delta)R\rfloor}, y_1^{j+\Delta}, x_1^{j-1})$ , having access to all encoded bits, a recent window of side information and all past source symbols,  $(b_1^{\lfloor(j+\Delta)R\rfloor}, y_j^{j+\Delta}, x_1^{j-1})$ , or having access to all encoded bits, all side information and all past decoding errors,  $(b_1^{\lfloor(j+\Delta)R\rfloor}, y_1^{j+\Delta}, \tilde{x}_1^{j-1})$ .

2) *Constructing a block code :* To encode a block of  $n$  source symbols, just run the rate  $R$  encoder  $\mathcal{E}$  and terminate with the encoder run using some  $\Delta$  random source symbols drawn according to the distribution of  $p_x$ . To decode the block, just use the delay  $\Delta$  rate  $R$  decoder  $\mathcal{D}^{\Delta,R}$  with feed-forward, and then further use the feedforward error signals to correct any mistakes that might have occurred. As a block coding system, this hypothetical system never makes an error from end to end. As shown in Figure 16, the data processing inequality implies:

*Lemma 4:* If  $n$  is the fixed block-length, and the block rate is  $R(1 + \frac{\Delta}{n})$ , then

$$H(\tilde{x}_1^n) \geq -(n + \Delta)R + nH(x|y) \quad (19)$$

*Proof:* See Appendix C.

3) *Lower bound the symbol-wise error probability :* Now suppose this block-code were to be run with the distribution  $q_{xy}$ , s.t.  $H(q_{x|y}) > (1 + \frac{\Delta}{n})R$ , from time 1 to  $n$ , and were to be run with the distribution  $p_{xy}$  from time  $n + 1$  to  $n + \Delta$ . Write the hybrid distribution as  $Q_{xy}$ . Then the block coding scheme constructed in the previous section would with probability very close to 1 make a block error. Moreover, many individual symbols would also be in error often:

*Lemma 5:* If the source and side-information is coming from  $q_{xy}$ , then there exists a  $\delta > 0$  so that for  $n$  large enough, there exists a number  $n_e$  and a sequence of symbol positions  $j_1 < j_2 < \dots < j_{n_e}$  satisfying:

- $n_e \geq \frac{H(q_{x|y}) - \frac{n+\Delta}{n}R}{2 \log_2 |\mathcal{X}| - (H(q_{x|y}) - \frac{n+\Delta}{n}R)} n$
- The probability of symbol errors made by the feed-forward decoder on symbol  $x_{j_i}$  is at least  $\delta$  when the joint source symbols are drawn according to  $q_{xy}$ .
- $\delta$  satisfies  $h_\delta + \delta \log_2(|\mathcal{X}| - 1) = \frac{1}{2}(H(q_{x|y}) - \frac{n+\Delta}{n}R)$  where  $h_\delta = -\delta \log_2 \delta - (1 - \delta) \log_2(1 - \delta)$ .

*Proof:* See Appendix D.

Pick  $j^* = j_{\frac{n_e}{2}}$  to pick a symbol position in the middle of the block that is subject to errors. Lemma 5 reveals that  $\min\{j^*, n - j^*\} \geq \frac{1}{2} \frac{(H(q_{x|y}) - \frac{n+\Delta}{n}R)}{2 \log_2 |\mathcal{X}| - (H(q_{x|y}) - \frac{n+\Delta}{n}R)} n$ , so if we fix  $\frac{\Delta}{n}$  and let  $n$  go to infinity, then  $\min\{j^*, n - j^*\}$  goes to infinity as well.

At this point, Lemma 2 implies that the decoder can ignore the side-information from the past. Define the “bad sequence” set  $E_{j^*}$  as the set of source and side-information sequence pairs so the type I delay  $\Delta$  rate  $R$  decoder makes a symbol error at  $j^*$ . To simplify notation, let  $\vec{x} = x_1^{j^*+\Delta}$ ,  $\bar{x} = x_1^{j^*-1}$ ,  $\bar{x} = x_{j^*}^{j^*+\Delta}$ ,  $\bar{y} = y_{j^*}^{j^*+\Delta}$  denote the entire source vector, the source prefix, and the suffixes for the source and side-information respectively. Define  $E_{j^*} = \{(\vec{x}, \bar{y}) | x_{j^*} \neq \mathcal{D}_{j^*}^{\Delta, R}(\mathcal{E}(\vec{x}), y_{j^*}^{j^*+\Delta}, \bar{x})\}$ .

Since the “bad sequence” set  $E_{j^*}$  only has future side-information in it, the probability of this set depends only on the marginals for  $x$  in the past and the joint distribution in the present and future. Consider a hybrid distribution where the joint source behaves according to  $Q_{xy}$  from time  $j^*$  to  $j^* + \Delta$  and the  $x$  source one behaves like it came from a distribution  $q_x$  from time 1 to  $j^* - 1$ . By Lemma 5,  $Q_{xy}(E_{j^*}) \geq \delta$ .

Define  $J = \min\{n, j^* + \Delta\}$  to deal with possible edge-effects<sup>7</sup> near the end of the block, and let  $\bar{\bar{x}} = x_{j^*}^J$ ,  $\bar{\bar{y}} = y_{j^*}^J$ . The empirical distribution of  $(\bar{\bar{x}}, \bar{\bar{y}})$  is written using shorthand  $r_{\bar{\bar{x}}, \bar{\bar{y}}}(x, y) = \frac{n_{x,y}(\bar{\bar{x}}, \bar{\bar{y}})}{\Delta+1}$  and similarly the empirical distribution of  $\bar{x}$  as  $r_{\bar{x}}(x) = \frac{n_x(\bar{x})}{j^*-1}$ .

Now, the strongly typical set can be defined

$$A_J^c(q_{xy}) = \left\{ \begin{array}{l} (\vec{x}, \bar{y}) \in \mathcal{X}^{j^*+\Delta} \times \mathcal{Y}^{\Delta+1} | \forall x, r_{\bar{x}}(x) \in (q_x(x) - \epsilon, q_x(x) + \epsilon) \\ \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, q_{xy}(x, y) > 0 : r_{\bar{\bar{x}}, \bar{\bar{y}}}(x, y) \in (q_{xy}(x, y) - \epsilon, q_{xy}(x, y) + \epsilon), \\ \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, q_{xy}(x, y) = 0 : r_{\bar{\bar{x}}, \bar{\bar{y}}}(x, y) = 0 \end{array} \right\}. \quad (20)$$

The conditions require that the prefix be  $q_x$ -typical and the suffix till  $J$  be  $q_{xy}$ -typical. What happens after  $J$  is not important.

This typical set is used to get a sequence of lemmas asserting that errors are common even when we restrict to the typical behavior of the  $q$  distribution, that the probability of  $q$ -typical joint realizations is least exponentially small under the true distribution, and that this means that the errors themselves must occur at least with exponentially small probability.

*Lemma 6:*  $Q_{xy}(E_{j^*} \cap A_J^c(q_{xy})) \geq \frac{\delta}{2}$  for large  $n$  and  $\Delta$ .

*Proof:* See Appendix E.

*Lemma 7:* For all  $\epsilon < \min_{x,y:p_{xy}(x,y)>0}\{p_{xy}(x,y)\}$ ,  $\forall (\vec{x}, \bar{y}) \in A_J^c(q_{xy})$ ,

$$\frac{p_{xy}(\vec{x}, \bar{y})}{Q_{xy}(\vec{x}, \bar{y})} \geq 2^{-(J-j^*+1)D(q_{xy}\|p_{xy}) - (j^*-1)D(q_x\|p_x) - JG\epsilon}$$

where  $G = \max\{|\mathcal{X}||\mathcal{Y}| + \sum_{x,y:p_{xy}(x,y)>0} \log_2(\frac{q_{xy}(x,y)}{p_{xy}(x,y)} + 1), |\mathcal{X}| + \sum_x \log_2(\frac{q_x(x)}{p_x(x)} + 1)\}$

*Proof:* See Appendix F.

*Lemma 8:* For all  $\epsilon < \min_{x,y}\{p_{xy}(x,y)\}$ , and large  $\Delta$ ,  $n$ :

$$p_{xy}(E_{j^*}) \geq \frac{\delta}{2} 2^{-(J-j^*+1)D(q_{xy}\|p_{xy}) - (j^*-1)D(q_x\|p_x) - JG\epsilon}$$

*Proof:* See Appendix G.

4) *Final details in proving Theorem 3:* Notice that as long as  $H(q_{x|y}) > \frac{n+\Delta}{n}R$ , we know  $\delta > 0$  by letting  $\epsilon$  go to 0, and having  $\Delta$  and  $n$  (and thus also  $J$ ) go to infinity proportionally. So  $\Pr[\hat{x}_{j^*}(j^* + \Delta) \neq x_{j^*}] = p_{xy}(E_{j^*}) \geq K 2^{-(J-j^*+1)D(q_{xy}\|p_{xy}) - (j^*-1)D(q_x\|p_x)}$ .

Notice that  $D(q_{xy}\|p_{xy}) \geq D(q_x\|p_x)$ . Since  $J = \min\{n, j^* + \Delta\}$ , for all possible  $j^* \in [1, n]$  we have for all  $n \geq \Delta$ :

$$\begin{aligned} (J - j^* + 1)D(q_{xy}\|p_{xy}) + (j^* - 1)D(q_x\|p_x) &\leq (\Delta + 1)D(q_{xy}\|p_{xy}) + (n - \Delta - 1)D(q_x\|p_x) \\ &\approx \Delta(D(q_{xy}\|p_{xy}) + \frac{n - \Delta}{\Delta}D(q_x\|p_x)). \end{aligned}$$

Meanwhile, for  $n < \Delta$ :

$$(J - j^* + 1)D(q_{xy}\|p_{xy}) + (j^* - 1)D(q_x\|p_x) \leq nD(q_{xy}\|p_{xy}) = \Delta(\frac{n}{\Delta}D(q_{xy}\|p_{xy})).$$

<sup>7</sup>These edge effects, although annoying, cannot be ignored since guaranteeing that  $\delta$  is small relative to  $n$  would come at the cost of less tight bounds in asymmetric cases. In this way, the situation is different from the argument given in [5] for channel coding without feedback.

Write  $\alpha = \frac{\Delta}{n}$ . The upper bound on the error exponent is the minimum of the above error exponents over all  $\alpha > 0$ .

$$E_{si}^u(R) = \min \left\{ \inf_{q_{xy}, \alpha \geq 1: H(q_{x|y}) > (1+\alpha)R} \left\{ \frac{1}{\alpha} D(q_{xy} \| p_{xy}) \right\}, \right. \\ \left. \inf_{q_{xy}, 1 \geq \alpha \geq 0: H(q_{x|y}) > (1+\alpha)R} \left\{ \frac{1-\alpha}{\alpha} D(q_x \| p_x) + D(q_{xy} \| p_{xy}) \right\} \right\}$$

This is the desired result. ■

The specialization of this result to uniform sources  $x$  and side information  $y = x \oplus s$  is straightforward and is covered in Appendix H. The key is to understand that when the joint source is symmetric, the marginal for  $q$  always agrees with the marginal for the original  $p$ .

## VI. CONCLUSIONS

This paper has shown that fixed-block-length and fixed-delay lossless source-coding behave very differently when decoder side-information is either present or absent at the encoder. While fixed-block-length systems do not usually gain substantially in reliability with encoder access to the side-information, fixed-delay systems can achieve very substantial gains in reliability. This means that if an application has a target for both end-to-end latency and probability of symbol error, then depriving the encoder of access to the side-information will come at the cost of higher required data rates.

The proof of achievability makes clear the connection to ideas of “effective bandwidth” and buffer-provisioning in the networking context (see e.g.[24]). The results here and in [5] can be considered a way to extend the spirit of those concepts to problems like source-coding without access to side-information and communication without feedback. Thinking about buffer-overflow is too narrow a perspective to generalize the idea of “how much extra rate is required beyond the minimum” but end-to-end delay provides a framework to understand this and thereby compare different approaches.

Thus, it is useful to view this paper as a companion to its sister paper [5] (treating channel-coding with and without feedback) in the fixed-delay context. Comparing both sets of results shows how feedback in channel coding is very much like encoder access to decoder side-information in lossless source coding. The main difference is that source coding performance is generally better at high rates while channel coding is better at low rates. The subtle aspect of the analogy is that lossless source-coding with encoder side-information behaves like channel-coding with feedback *for channels with strictly positive zero-error capacity*.

- For generic symmetric channels with  $C_{0,f} > 0$ , the fixed-block-length reliability function is known perfectly with feedback and jumps abruptly to  $\infty$  at  $C_{0,f}$  and approaches zero quadratically at  $C$ .  
For generic sources, the fixed-block-length reliability function is known perfectly with encoder side-information and jumps abruptly to  $\infty$  at  $\log_2 |\mathcal{X}|$  and approaches zero quadratically at  $H_{x|y}$ .
- For generic symmetric channels with  $C_{0,f} > 0$ , the fixed-delay reliability with feedback tends smoothly to  $\infty$  at  $C_{0,f}$  and approaches zero linearly at  $C$ .  
For general sources with encoder access to side-information, the fixed-delay reliability function tends smoothly to  $\infty$  at  $\log_2 |\mathcal{X}|$  and approaches zero linearly at  $H_{x|y}$ .
- For generic symmetric channels with  $C_{0,f} > 0$ , an asymptotically optimal fixed-delay code with feedback can be constructed using a queue fed at fixed rate followed by a fixed-to-variable channel code.  
For generic sources with encoder access to side-information, an asymptotically optimal fixed-delay code can be constructed using a fixed-to-variable source code followed by a queue drained at fixed rate.

In both cases, the non-ignorant encoders can help deliver substantially lower end-to-end delays. In addition, in both cases there is a gap between the achievable regions and converses for fixed delay reliability for asymmetric cases when considering ignorant encoders. In addition to closing this gap, many natural problems remain to be explored: joint source-channel coding [25], lossy coding [26], as well as extending the upper-bound techniques here to truly multi-terminal settings with distributed encoders.

APPENDIX A  
PROOF OF LEMMA 1

In the large deviation theory literature, limit superior and limit inferior are widely used while calculating the asymptotic properties of the rate functions [27]. However it is sometimes more convenient to use the following equivalent  $\epsilon - K$  conditions since

$$a \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log_2 P_n \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log_2 P_n \leq b$$

iff for all  $\epsilon > 0$ , there exists  $K < \infty$ , such that for all  $n$ :  $K2^{n(a-\epsilon)} \leq P_n \leq K2^{n(b+\epsilon)}$ . The equivalence is obvious from the definitions of limit superior and limit inferior [27].

*Proof:* By Cramér's theorem[27], for all  $\epsilon_1 > 0$  there exists  $K_1$ , such that:

$$\Pr\left(\sum_{i=1}^n l(\vec{x}_i, \vec{y}_i) > nNr\right) = \Pr\left(\frac{1}{n} \sum_{i=1}^n l(\vec{x}_i, \vec{y}_i) > Nr\right) \leq K_1 2^{-n(\inf_{z > Nr} I(z) - \epsilon_1)} \quad (21)$$

where the rate function  $I(z)$  is [27]:

$$I(z) = \sup_{\rho \in \mathcal{R}} \{ \rho z - \log_2 \left( \sum_{(\vec{x}, \vec{y}) \in \mathcal{X}^N \times \mathcal{Y}^N} p_{xy}(\vec{x}, \vec{y}) 2^{\rho l(\vec{x}, \vec{y})} \right) \} \quad (22)$$

$$\text{Write } I(z, \rho) = \rho z - \log_2 \left( \sum_{(\vec{x}, \vec{y}) \in \mathcal{X}^N \times \mathcal{Y}^N} p_{xy}(\vec{x}, \vec{y}) 2^{\rho l(\vec{x}, \vec{y})} \right)$$

Notice that the Hölder inequality implies that for all  $\rho_1, \rho_2$  and for all  $\theta \in (0, 1)$ :

$$\begin{aligned} \left( \sum_i p_i 2^{\rho_1 l_i} \right)^\theta \left( \sum_i p_i 2^{\rho_2 l_i} \right)^{(1-\theta)} &\geq \sum_i (p_i^\theta 2^{\theta \rho_1 l_i}) (p_i^{1-\theta} 2^{(1-\theta) \rho_2 l_i}) \\ &= \sum_i p_i 2^{(\theta \rho_1 + (1-\theta) \rho_2) l_i} \end{aligned}$$

This shows that  $\log_2 \left( \sum_{(\vec{x}, \vec{y}) \in \mathcal{X}^N \times \mathcal{Y}^N} p_{xy}(\vec{x}, \vec{y}) 2^{\rho l(\vec{x}, \vec{y})} \right)$  is a convex  $\cup$  function of  $\rho$  and thus  $I(z, \rho)$  is a concave  $\cap$  function of  $\rho$  for fixed  $z$ . Clearly  $I(z, 0) = 0$ . Consider  $z > Nr > NH(x|y)$ . For large  $N$ ,

$$\frac{\partial I(z, \rho)}{\partial \rho} \Big|_{\rho=0} = z - \sum_{(\vec{x}, \vec{y}) \in \mathcal{X}^N \times \mathcal{Y}^N} p_{xy}(\vec{x}, \vec{y}) l(\vec{x}, \vec{y}) \geq 0$$

since the average codeword length is essentially  $NH(x|y)$ . Thus  $I(z, \rho) < 0$  as long as  $z > Nr$  and  $\forall \rho < 0$ . This means that the  $\rho$  to maximize  $I(z, \rho)$  is positive. So from now on, it is safe to assume  $\rho \geq 0$ . This implies that  $I(z)$  is monotonically increasing with  $z$  and it is obvious that  $I(z)$  is continuous. Thus

$$\inf_{z > Nr} I(z) = I(Nr). \quad (23)$$

Using the upper bound on  $l(\vec{x}, \vec{y})$  in (14):

$$\begin{aligned} \log_2 \left( \sum_{(\vec{x}, \vec{y}) \in \mathcal{X}^N \times \mathcal{Y}^N} p_{xy}(\vec{x}, \vec{y}) 2^{\rho l(\vec{x}, \vec{y})} \right) &\leq \log_2 \left( \sum_{q_{xy} \in \mathcal{T}^N} 2^{-ND(q_{xy} \| p_{xy})} 2^{\rho(\epsilon_N + NH(q_{x|y}))} \right) \\ &\leq 2^{N\epsilon_N} 2^{-N \min_q \{ D(q_{xy} \| p_{xy}) - \rho H(q_{x|y}) - \rho \epsilon_N \}} \\ &= N \left( - \min_q \{ D(q_{xy} \| p_{xy}) - \rho H(q_{x|y}) - \rho \epsilon_N \} + \epsilon_N \right) \end{aligned}$$

where  $0 < \epsilon_N \leq \frac{2+|\mathcal{X}||\mathcal{Y}|\log_2(N+1)}{N}$  goes to 0 as  $N$  goes to infinity and  $\mathcal{T}^N$  is the set of all joint types of  $\mathcal{X}^N \times \mathcal{Y}^N$ .

Substitute the above inequalities into  $I(Nr)$  defined in (22):

$$I(Nr) \geq N \left( \sup_{\rho \geq 0} \left\{ \min_q \left\{ \rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy} \| p_{xy}) \right\} - \epsilon_N \right) \right) \quad (24)$$

The next task is to show that  $I(Nr) \geq N(E_{ei,b}(r) + \epsilon)$  where  $\epsilon$  goes to 0 as  $N$  goes to infinity. This can be proved by the tedious but direct Lagrange multiplier method used in [17]. Instead, the proof here is based on the existence of a saddle point. Define

$$f(q, \rho) = \rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy} \| p_{xy}).$$

Clearly for fixed  $q$ ,  $f(q, \rho)$  is a linear function of  $\rho$ , and thus concave. In addition, for fixed  $\rho \geq 0$ ,  $f(q, \rho)$  is a convex  $\cup$  function of  $q$ , because both  $-H(q_{x|y})$  and  $D(q_{xy} \| p_{xy})$  are convex  $\cup$  on  $q_{xy}$ . Define  $g(u) \triangleq \min_q \sup_{\rho \geq 0} (f(q, \rho) + \rho u)$ . It is enough to show that  $g(u)$  is finite in the neighborhood of  $u = 0$  to establish the existence of the saddle point [28].

$$\begin{aligned} g(u) & \stackrel{(a)}{=} \min_q \sup_{\rho \geq 0} f(q, \rho) + \rho u \\ & \stackrel{(b)}{=} \min_q \sup_{\rho \geq 0} \rho(r - H(q_{x|y}) - \epsilon_N + u) + D(q_{xy} \| p_{xy}) \\ & \leq (c) \min_{q: H(q_{x|y}) \geq r - \epsilon_N + u} \sup_{\rho \geq 0} \rho(r - H(q_{x|y}) - \epsilon_N + u) + D(q_{xy} \| p_{xy}) \\ & \leq (d) \min_{q: H(q_{x|y}) \geq r - \epsilon_N + u} D(q_{xy} \| p_{xy}) \\ & \stackrel{(e)}{<} \infty \end{aligned} \tag{25}$$

(a), (b) are from the definitions. (c) is true because  $H(p_{x|y}) < r < \log_2 |\mathcal{X}|$  and thus for very small  $\epsilon_N$  and  $u$ ,  $H(p_{x|y}) < r - \epsilon_N + u < \log_2 |\mathcal{X}|$ . Consequently, there exists a distribution  $q$  so that  $H(q_{x|y}) \geq r - \epsilon_N + u$ . (d) holds because  $H(q_{x|y}) \geq r - \epsilon_N + u$  and  $\rho \geq 0$ . (e) is true because we assumed without loss of generality that the marginal  $p_x(x) > 0$  for all  $x \in \mathcal{X}$  together with the fact that  $r - \epsilon_N + u < \log_2 |\mathcal{X}|$ . The finiteness implies the existence of the saddle point of  $f(q, \rho)$ .

$$\sup_{\rho \geq 0} \{ \min_q f(q, \rho) \} = \min_q \{ \sup_{\rho \geq 0} f(q, \rho) \} \tag{26}$$

Note that if  $H(q_{x|y}) < r + \epsilon_N$ , then  $\rho$  can be chosen to be arbitrarily large to make  $\rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy} \| p_{xy})$  arbitrarily large. Thus the  $q$  to minimize  $\sup_{\rho} \rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy} \| p_{xy})$  satisfies  $r - H(q_{x|y}) - \epsilon_N \geq 0$ . Thus

$$\begin{aligned} \min_q \{ \sup_{\rho \geq 0} \rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy} \| p_{xy}) \} & \stackrel{(a)}{=} \min_{q: H(q_{x|y}) \geq r - \epsilon_N} \sup_{\rho \geq 0} \{ \rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy} \| p_{xy}) \} \\ & \stackrel{(b)}{=} \min_{q: H(q_{x|y}) \geq r - \epsilon_N} \{ D(q_{xy} \| p_{xy}) \} \\ & \stackrel{(c)}{=} E_{ei,b}(r - \epsilon_N). \end{aligned} \tag{27}$$

(a) follows from the argument above. (b) is true because  $r - H(q_{x|y}) - \epsilon_N \leq 0$  and  $\rho \geq 0$  and thus  $\rho = 0$  maximizes  $\rho(r - H(q_{x|y}) - \epsilon_N)$ . (c) is true by definition. Combining (24) (26) and (27), letting  $N$  be sufficiently big implies that  $\epsilon_N$  is sufficiently small. Noticing that  $E_{ei}(r)$  is continuous on  $r$ , we get the the desired bound in (15).  $\square$

## APPENDIX B PARAMETRIZATION OF $E_{ei}(R)$

We need the definition of tilted distributions for a joint distribution  $p_{xy}$  from [17].

*Definition 5:*  $x - y$  tilted distribution of  $p_{xy}$ :  $\bar{p}_{xy}^\rho$ , for all  $\rho \in [-1, +\infty)$

$$\bar{p}_{xy}^\rho(x, y) = \frac{[\sum_s p_{xy}(s, y)^{\frac{1}{1+\rho}}]^{1+\rho}}{\sum_t [\sum_s p_{xy}(s, t)^{\frac{1}{1+\rho}}]^{1+\rho}} \times \frac{p_{xy}(x, y)^{\frac{1}{1+\rho}}}{\sum_s p_{xy}(s, y)^{\frac{1}{1+\rho}}} \tag{28}$$

Write the conditional entropy of  $x$  given  $y$  for this tilted distribution as  $H(\bar{p}_{x|y}^\rho)$ . An important fact as shown in Lemma 17 of [17] is that  $\frac{\partial E_0(\rho)}{\partial \rho} = H(\bar{p}_{x|y}^\rho)$ , also  $H(\bar{p}_{x|y}^\rho)|_{\rho=0} = H(p_{x|y})$ ,  $H(\bar{p}_{x|y}^\rho)|_{\rho=+\infty} = \log_2(M(p_{xy}))$ . where  $M(p_{xy}) = |\max_{y \in \mathcal{Y}} \{x \in \mathcal{X} : p_{xy}(x, y) > 0\}|$

We first show that  $\frac{E_0(\rho)}{\rho}$  is in general monotonically increasing for  $\rho \in [0, \infty)$ :

$$\begin{aligned}
\frac{\partial \frac{E_0(\rho)}{\rho}}{\partial \rho} & \stackrel{(a)}{=} \frac{\rho \frac{\partial E_0(\rho)}{\partial \rho} - E_0(\rho)}{\rho^2} \\
& \stackrel{(b)}{=} \frac{\rho H(\bar{p}_{x|y}^\rho) - E_0(\rho)}{\rho^2} \\
& \stackrel{(c)}{=} \frac{D(\bar{p}_{xy}^\rho \| p_{xy})}{\rho^2} \\
& \stackrel{(d)}{>} 0
\end{aligned}$$

(a) is obvious and (b) is from Lemma 17 in [17]. (c) is from Lemma 15 in [17]. (d) is true unless the source  $x$  is conditionally uniform given side information  $y$ . For the trivial case conditionally uniform case where  $p_{x|y}(x|y) = \frac{1}{M(p_{xy})}$  on those letters  $x$  for which it is nonzero, both the fixed-block-length error exponent  $E_{si,b}^u(R)$  and the delay error exponent  $E_{ei}(R)$  are either 0 when  $R < \log_2(M(p_{xy}))$  or  $\infty$  when  $R > \log_2(M(p_{xy}))$ .

With the above observations, we know that for all  $R \in [H(p_{x|y}), \log_2 M(p_{xy})]$ , there exists a unique  $\rho^* \geq 0$ , s.t.  $R = \frac{E_0(\rho^*)}{\rho^*}$  or equivalently  $\rho^* R = E_0(\rho^*)$ . In order to show (11), it remains to show that  $E_{ei}(R) = E_0(\rho^*)$ .

From the definition of  $E_{ei}(R)$  in (10) and the definition of  $E_{si,b}^u(R)$  in (5), we have:

$$\begin{aligned}
E_{ei}(R) & = \inf_{\alpha > 0} \frac{1}{\alpha} E_{si,b}^u((\alpha + 1)R) \\
& = \inf_{\alpha > 0} \left\{ \sup_{\rho \geq 0} \frac{\rho(\alpha + 1)R - E_0(\rho)}{\alpha} \right\} \\
& \geq \sup_{\rho \geq 0} \inf_{\alpha > 0} \rho R + \frac{\rho R - E_0(\rho)}{\alpha} \\
& \geq \inf_{\alpha > 0} \rho^* R + \frac{\rho^* R - E_0(\rho^*)}{\alpha} \\
& = \rho^* R.
\end{aligned} \tag{29}$$

Now show that  $E_{ei}(R) \leq \rho^* R$  by writing  $\rho(R)$  as the parameter  $\rho$  that maximizes  $\rho R - E_0(\rho)$ . From the convexity of  $E_0(\rho)$  for  $\rho \in [0, \infty)$  and the fact that  $R \in [H(p_{x|y}), \log_2 M(p_{xy})]$ , we know that  $\rho(R)$  is the unique positive real number s.t.  $R = \frac{\partial E_0(\rho)}{\partial \rho} |_{\rho=\rho(R)} = H(\bar{p}_{x|y}^\rho) |_{\rho=\rho(R)}$ .  $\rho R - E_0(\rho)$  is a concave  $\cap$  function of  $\rho$  and  $\rho R - E_0(\rho) |_{\rho=0} = 0$ , hence  $\rho(R) \leq \rho^*$  where  $\rho(R)$  is the maximal point and  $\rho^*$  is the zero point of  $\rho R - E_0(\rho)$ . This is illustrated in Figure 17.

Because  $R \in [H(p_{x|y}), \log_2 M(p_{xy})]$  and  $\rho(R) < \rho^*$ , there exists  $R' > R$ , s.t.  $\rho^* = \rho(R')$ , i.e.  $\rho^*$  maximizes  $\rho R' - E_0(\rho)$ . Now let  $\alpha^* = \frac{R'}{R} - 1$  which is positive because  $R' > R$ . That is  $\rho^*$  maximizes  $\rho R' - E_0(\rho) = \rho(1 + \alpha^*)R - E_0(\rho)$ . Plugging this in, gives:

$$\begin{aligned}
E_{ei}(R) & = \inf_{\alpha > 0} \left\{ \sup_{\rho \geq 0} \frac{\rho(\alpha + 1)R - E_0(\rho)}{\alpha} \right\} \\
& \leq \sup_{\rho \geq 0} \frac{\rho(1 + \alpha^*)R - E_0(\rho)}{\alpha^*} \\
& = \frac{\rho^*(1 + \alpha^*)R - E_0(\rho^*)}{\alpha^*} \\
& = \rho^* R = E_0(\rho^*).
\end{aligned} \tag{30}$$

Finally, to get the slope in the vicinity of the conditional entropy, just expand  $E_0(\rho)$  around  $\rho = 0$  using a Taylor series. The constant term is zero and Lemma 17 of [17] reveals that the first order term is the conditional entropy itself. The slope  $\frac{\partial E}{\partial \rho} / \frac{\partial R}{\partial \rho}$  evaluated at  $\rho = 0$  is clearly the first-order term in the Taylor series divided by the second order term, giving the desired result. The second derivative of  $E_0(\rho)$  is only zero when  $D(\bar{p}_{xy}^0 \| p_{xy}) = 0$  which implies that  $p_{xy}$  is itself conditionally uniform, resulting in the claimed infinite error exponents.

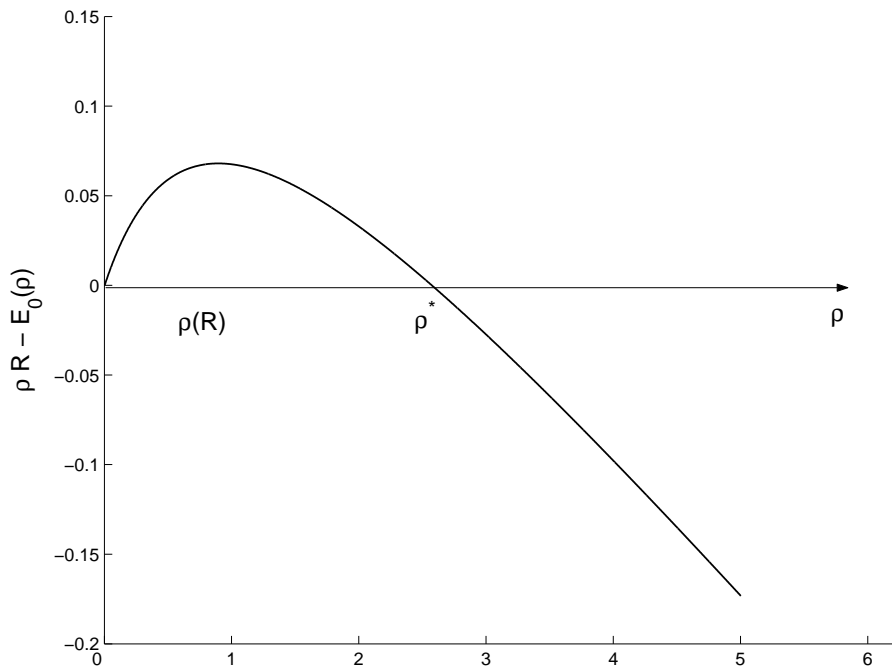


Fig. 17. Plot of  $\rho R - E_0(\rho)$ . The maximizing  $\rho(R) < \rho^*$ , the point where this crosses zero.  $R$  is fixed.

#### APPENDIX C PROOF OF LEMMA 4

$$\begin{aligned}
nH(x) & \stackrel{(a)}{=} H(x_1^n) \\
& = I(x_1^n; x_1^n) \\
& \stackrel{(b)}{=} I(x_1^n; \tilde{x}_1^n, \mathbf{b}_1^{\lfloor (n+\Delta)R \rfloor}, y_1^{n+\Delta}) \\
& \stackrel{(c)}{=} I(x_1^n; y_1^{n+\Delta}) + I(x_1^n; \tilde{x}_1^n | y_1^{n+\Delta}) + I(x_1^n; \mathbf{b}_1^{\lfloor (n+\Delta)R \rfloor} | y_1^{n+\Delta}, \tilde{x}_1^n) \\
& \stackrel{(d)}{\leq} nI(x, y) + H(\tilde{x}_1^n) + H(\mathbf{b}_1^{\lfloor (n+\Delta)R \rfloor}) \\
& \leq nH(x) - nH(x|y) + H(\tilde{x}_1^n) + (n + \Delta)R
\end{aligned}$$

(a) is true because the source is iid. (b) is true because of the data processing inequality considering the following Markov chain:  $x_1^n - (\tilde{x}_1^n, \mathbf{b}_1^{\lfloor (n+\Delta)R \rfloor}, y_1^n) - x_1^n$ , thus  $I(x_1^n; x_1^n) \leq I(x_1^n; \tilde{x}_1^n, \mathbf{b}_1^{\lfloor (n+\Delta)R \rfloor}, y_1^{n+\Delta})$ . Furthermore,  $I(x_1^n; x_1^n) = H(x_1^n) \geq I(x_1^n; \tilde{x}_1^n, \mathbf{b}_1^{\lfloor (n+\Delta)R \rfloor}, y_1^{n+\Delta})$ . Combining the two inequalities gives (b). (c) is the chain rule for mutual information. In (d), first notice that  $(x, y)$  are iid across time and thus  $I(x_1^n; y_1^{n+\Delta}) = I(x_1^n; y_1^n) = nI(x, y)$ . Second, the entropy of a random variable is never less than the mutual information of that random variable with another one, conditioned on another random variable or not.  $\square$

#### APPENDIX D PROOF OF LEMMA 5

Lemma 4 implies:

$$\sum_{i=1}^n H(\tilde{x}_i) \geq H(\tilde{x}_1^n) \geq -(n + \Delta)R + nH(q_{x|y}) \tag{31}$$

The average entropy per source symbol for  $\tilde{x}$  is at least  $H(q_{x|y}) - \frac{n+\Delta}{n}R$ . Now suppose that  $H(\tilde{x}_i) \geq \frac{1}{2}(H(q_{x|y}) - \frac{n+\Delta}{n}R)$  for  $n_e$  symbol positions  $1 \leq j_1 < j_2 < \dots < j_A \leq n$ . By noticing that  $H(\tilde{x}_i) \leq \log_2 |\mathcal{X}|$ , we have

$$\sum_{i=1}^n H(\tilde{x}_i) \leq n_e \log_2 |\mathcal{X}| + (n - n_e) \frac{1}{2} (H(q_{x|y}) - \frac{n+\Delta}{n}R)$$

Combining this with (31) gives:

$$n_e \geq \frac{(H(q_{x|y}) - \frac{n+\Delta}{n}R)}{2 \log_2 |\mathcal{X}| - (H(q_{x|y}) - \frac{n+\Delta}{n}R)} n \quad (32)$$

where  $2 \log_2 |\mathcal{X}| - (H(q_{x|y}) - \frac{n+\Delta}{n}R) \geq 2 \log_2 |\mathcal{X}| - H(q_{x|y}) \geq 2 \log_2 |\mathcal{X}| - \log_2 |\mathcal{X}| > 0$ .

For each of the  $j$ , the individual entropy  $H(\tilde{x}_j) \geq \frac{1}{2}(H(q_{x|y}) - \frac{n+\Delta}{n}R)$ . By the monotonicity of the binary entropy function,  $\Pr(\tilde{x}_j \neq x_0) = \Pr(x_j \neq \tilde{x}_j) \geq \delta$ .  $\square$

#### APPENDIX E PROOF OF LEMMA 6

If we fix  $\frac{\Delta}{n}$  and let  $n$  go to infinity, then by definition  $J = \min\{n, j^* + \Delta\}$  goes to infinity as well. By Lemma 13.6.1 in [19], it is known that  $\forall \epsilon > 0$ , since  $J - j^*$  and  $j^*$  both getting large with  $n$ , that  $Q_{xy}(A_J^\epsilon(q_{xy}))^C \leq \frac{\delta}{2}$ . By Lemma 5,  $Q_{xy}(E_{j^*}) \geq \delta$ . So

$$Q_{xy}(E_{j^*} \cap A_J^\epsilon(q_{xy})) \geq Q_{xy}(E_{j^*}) - Q_{xy}(A_J^\epsilon(q_{xy}))^C \geq \frac{\delta}{2}$$

$\square$

#### APPENDIX F PROOF OF LEMMA 7

For  $(\vec{x}, \vec{y}) \in A_J^\epsilon(q_{xy})$ , by the definition of the strongly typical set, it can be easily shown by algebra that  $D(r_{\vec{x}, \vec{y}} \| p_{xy}) \leq D(q_{xy} \| p_{xy}) + G\epsilon$  and  $D(r_{\vec{x}} \| p_x) \leq D(q_x \| p_x) + G\epsilon$ . So

$$\begin{aligned} \frac{p_{xy}(\vec{x}, \vec{y})}{Q_{xy}(\vec{x}, \vec{y})} &= \frac{p_{xy}(\vec{x}) p_{xy}(\vec{\bar{x}}, \vec{\bar{x}}) p_{xy}(x_{J+1}^{j^*+\Delta}, y_{J+1}^{j^*+\Delta})}{q_{xy}(\vec{x}) q_{xy}(\vec{\bar{x}}, \vec{\bar{y}}) p_{xy}(x_{J+1}^{j^*+\Delta}, y_{J+1}^{j^*+\Delta})} \\ &= \frac{2^{-(J-j^*+1)(D(r_{\vec{x}, \vec{y}} \| p_{xy}) + H(r_{\vec{x}, \vec{y}}))} 2^{-(j^*-1)(D(r_{\vec{x}} \| p_x) + H(r_{\vec{x}}))}}{2^{-(J-j^*+1)(D(r_{\vec{x}, \vec{y}} \| q_{xy}) + H(r_{\vec{x}, \vec{y}}))} 2^{-(j^*-1)(D(r_{\vec{x}} \| q_x) + H(r_{\vec{x}}))}} \\ &\stackrel{(a)}{\geq} \frac{2^{-(J-j^*+1)(D(q_{xy} \| p_{xy}) + G\epsilon) - (j^*-1)(D(q_x \| p_x) + G\epsilon)}}{2^{-(J-j^*+1)D(q_{xy} \| p_{xy}) - (j^*-1)D(q_x \| p_x) - JG\epsilon}} \end{aligned}$$

where (a) is true by (12.60) in [19].  $\square$

#### APPENDIX G PROOF OF LEMMA 8

Combining Lemmas 6 and 7:

$$\begin{aligned} p_{xy}(E_{j^*}) &\geq p_{xy}(E_{j^*} \cap A_J^\epsilon(q_{xy})) \\ &\geq q_{xy}(E_{j^*} \cap A_J^\epsilon(q_{xy})) 2^{-(J-j^*+1)D(q_{xy} \| p_{xy}) - (j^*-1)D(q_x \| p_x) - JG\epsilon} \\ &\geq \frac{\delta}{2} 2^{-(J-j^*+1)D(q_{xy} \| p_{xy}) - (j^*-1)D(q_x \| p_x) - JG\epsilon} \end{aligned}$$

$\square$

APPENDIX H  
PROOF OF COROLLARY 1

Theorem 3 asserts that

$$\begin{aligned}
E_{s_i}(R) &\leq \left\{ \inf_{q_{xy}, \alpha \geq 1: H(q_{x|y}) > (1+\alpha)R} \left\{ \frac{1}{\alpha} D(q_{xy} \| p_{xy}) \right\}, \right. \\
&\quad \left. \inf_{q_{xy}, 1 \geq \alpha \geq 0: H(q_{x|y}) > (1+\alpha)R} \left\{ \frac{1-\alpha}{\alpha} D(q_x \| p_x) + D(q_{xy} \| p_{xy}) \right\} \right\} \\
&\leq \inf_{q_{xy}, 1 \geq \alpha \geq 0: H(q_{x|y}) > (1+\alpha)R} \left\{ \frac{1-\alpha}{\alpha} D(q_x \| p_x) + D(q_{xy} \| p_{xy}) \right\} \\
&\leq \inf_{q_{xy}, 1 \geq \alpha \geq 0: H(q_{x|y}) > (1+\alpha)R, q_x = p_x} \left\{ \frac{1-\alpha}{\alpha} D(q_x \| p_x) + D(q_{xy} \| p_{xy}) \right\} \\
&= \inf_{q_{xy}, 1 \geq \alpha \geq 0: H(q_{x|y}) > (1+\alpha)R, q_x = p_x} \left\{ D(q_{xy} \| p_{xy}) \right\} \\
&= \inf_{q_{xy}: H(q_{x|y}) > R, q_x = p_x} \left\{ D(q_{xy} \| p_{xy}) \right\} \tag{33}
\end{aligned}$$

The next step is to show that (33) is indeed  $E_{s,b}(R, p_s)$  for uniform sources  $x$  and symmetric side information  $y$ , where  $x = y \oplus s$ .

$$\begin{aligned}
\inf_{q_{xy}: H(q_{x|y}) > R, q_x = p_x} \left\{ D(q_{xy} \| p_{xy}) \right\} &= (a) \inf_{q_{xy}: H(q_{x|y}) > R} \left\{ D(q_{xy} \| p_{xy}) \right\} \\
&= (b) \max_{\rho \geq 0} \rho R - E_0(\rho) \\
&= (c) \max_{\rho \geq 0} \rho R - (1 + \rho) \log \left[ \sum_s p_s(s)^{\frac{1}{1+\rho}} \right] \\
&= (d) E_{s,b}(R, p_s) \tag{34}
\end{aligned}$$

(b) follows from (5) in Theorem 1, (c) follows since (6) can be simplified for this case:

$$\begin{aligned}
E_0(\rho) &= \log_2 \sum_y \left( \sum_x p_{xy}(x, y)^{\frac{1}{1+\rho}} \right)^{(1+\rho)} \\
&= \log_2 \sum_y \left( \sum_x (p_y(y) p_{x|y}(x|y))^{\frac{1}{1+\rho}} \right)^{(1+\rho)} \\
&= \log_2 \sum_y p_y(y) \left( \sum_x (p_{x|y}(x|y))^{\frac{1}{1+\rho}} \right)^{(1+\rho)} \\
&= \log_2 \sum_y p_y(y) \left( \sum_s (p_s(s))^{\frac{1}{1+\rho}} \right)^{(1+\rho)} \\
&= \log_2 \left( \sum_s (p_s(s))^{\frac{1}{1+\rho}} \right)^{(1+\rho)} \\
&= (1 + \rho) \log_2 \left( \sum_s p_s(s)^{\frac{1}{1+\rho}} \right) \tag{35}
\end{aligned}$$

where this clearly matches from (9) to give us (d).

Thus, for uniform source  $x$  and side information  $y = x \ominus s$ , the distribution  $q_{xy}$  that minimizes the RHS of (34) is also marginally uniform on  $x$  since all that needs to tilt is the distribution for  $s$ . Hence the constraint on the marginal  $q_x = p_x$  is redundant and (a) is true.  $\square$

REFERENCES

- [1] C. E. Shannon, "The zero error capacity of a noisy channel," *IEEE Trans. Inform. Theory*, vol. 2, no. 3, pp. 8–19, Sept. 1956.
- [2] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, July 1973.
- [3] R. L. Dobrushin, "An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback," *Problemy Kibernetiki*, vol. 8, pp. 161–168, 1962.

- [4] E. R. Berlekamp, "Block coding with noiseless feedback," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1964.
- [5] A. Sahai, "Why block-length and delay behave differently if feedback is present," *IEEE Trans. Inform. Theory*, Submitted. [Online]. Available: <http://www.eecs.berkeley.edu/~sahai/Papers/FocusingBound.pdf>
- [6] R. G. Gallager, "Source coding with side information and universal coding," Massachusetts Institute of Technology, Tech. Rep. LIDS-P-937, 1976.
- [7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.
- [8] T. M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with two-sided side information," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1629–1638, June 2002.
- [9] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extensions to the side information case," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1181–1203, May 2003.
- [10] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *IEEE Trans. Inform. Theory*, vol. 49, pp. 626–643, Mar. 2003.
- [11] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, pp. 71–83, 2005.
- [12] M. Johnson, P. Ishwar, V. Prabhakaran, D. Schonberg, and K. Ramchandran, "On compressing encrypted data," *IEEE Trans. Signal Processing*, vol. 52, no. 10, pp. 2992–3006, 2004.
- [13] C. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.
- [14] M. V. Burnashev, "Data transmission over a discrete channel with feedback, random transmission time," *Problemy Peredachi Informatsii*, vol. 12, no. 4, pp. 10–30, Oct./Dec. 1976.
- [15] M. S. Pinsker, "Bounds on the probability and of the number of correctable errors for nonblock codes," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 44–55, Oct./Dec. 1967.
- [16] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record*, vol. 7, no. 4, pp. 142–163, 1959.
- [17] C. Chang, S. Draper, and A. Sahai, "Lossless coding for distributed streaming sources," *IEEE Trans. Inform. Theory*, submitted.
- [18] H. Palaiyanur and A. Sahai, "Sequential decoding using side-information," *IEEE Trans. Inform. Theory*, submitted.
- [19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [20] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Trans. Inform. Theory*, vol. 14, pp. 490–501, 1968.
- [21] N. Merhav, "Universal coding with minimum probability of codeword lengthoverflow," *IEEE Trans. Inform. Theory*, vol. 37, pp. 556–563, May 1991.
- [22] R. Durrett, *Probability: Theory and Examples*. Belmont, CA: Brooks/Cole, 2005.
- [23] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.
- [24] C. Chang and J. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, pp. 1091–1114, Aug. 1995.
- [25] C. Chang and A. Sahai, "Error exponents with delay for joint source channel coding," in *Forty-fourth Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sept. 2006.
- [26] —, "Delay-constrained source coding for a peak distortion measure," in *Proceedings of the 2007 IEEE Symposium on Information Theory*, Nice, France, July 2007.
- [27] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 1998.
- [28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.